JOSEPH D. LEONE

Recombinant DNA

SECOND EDITION

James D. Watson
COLD SPRING HARBOR LABORATORY

Michael Gilman
COLD SPRING HARBOR LABORATORY

Jan Witkowski
BANBURY CENTER, COLD SPRING HARBOR LABORATORY

Mark Zoller
GENENTECH, INC.

SCIENTIFIC
AMERICAN
BOOKS

Distributed by
W. H. Freeman and Company
New York

The cover illustration, by Marvin Mattelson, symbolizes some of the elements of this book. The DNA double helix is, of course, central to the book, as it is to the cover illustration. The blocks are double-stranded DNA fragments synthesized by the polymerase chain reaction, a technique that has revolutionized the way molecular genetics experiments are done. The number of fragments doubles repeatedly, going off into the distance (see Chapter 6). The coat colors of the mice running down the helix (in the same direction but with opposite polarity!), are changing from albino to chimeric, then chimeric to agouti. These coat color changes show mice in which genetic engineering has been used to knock out a specific gene. The experiment is shown more realistically in Figure 14-9.

Library of Congress Cataloging-in-Publication Data.

Recombinant DNA/James D. Watson... [et al.]. — 2nd ed.
p. cm.
Includes bibliographical references and index.
ISBN 0-7167-1994-0. — ISBN 0-7167-2282-8 (pbk.)
1. Recombinant DNA. I. Watson, James D., 1928–
QH442.R37 1992
574.87'3282—dc20 91-38483
CIP

Printed in the United States of America

Scientific American Books is a subsidiary of Scientific American, Inc.

Distributed by W. H. Freeman and Company, 41 Madison Avenue, New York, New York 10010

4 5 6 7 8 9 0 RRD 9 9 8 7 6 5 4

"Wobble" Frequently Permits Single tRNA Species to Recognize Multiple Codons

Initially it seemed highly probable that anticodons bind codons by means of three A—U and/or G—C hydrogen bonds that are identical to the A—T and G—C bonds that bind the two strands of the double helix together. But soon experimental results began to show that single tRNA species can bind to, say, both UUU and UUC codons. This suggested to Francis Crick that although the first two bases in a codon always pair in a DNA-like fashion, the pairing in the third position is less restrictive, so that nonstandard base pairing (wobble) is allowed for the third bases. Additional types of pairing became possible, opening up the possibility that there need not always be a distinct tRNA species for each of the 61 codons corresponding to amino acids. To date, the most complete data come from yeasts that have been shown, from a combination of genetic and DNA sequence data, to contain approximately 45 tRNA species. Thus, many yeast tRNA molecules have to recognize more than one codon.

Great variation exists in the relative amounts of particular tRNA species that are present in a given cell. In part, this variation reflects differences in the abundance of the amino acids the tRNAs specify. For example, the amino acids methionine and tryptophan occur relatively rarely in most proteins, and comparatively small amounts of their respective tRNAs are present. Moreover, when more than one tRNA form exists for a given amino acid, these different tRNA forms tend not to be present in equal amounts. This suggested that the more numerous tRNAs recognize the more commonly used codons for a given amino acid. This supposition was proved correct when methods for determining the precise nucleotide sequences of genes became available (Chapter 5). The rate at which an mRNA message is translated into its corresponding polypeptide chain thus may be controlled in part by whether it contains codons that are recognized by the more commonly available tRNA forms.

How Universal Is the Genetic Code?

Virtually all the experiments used to decipher the genetic code employed ribosomes and tRNA molecules from *E. coli*. It could thus be asked whether mRNA molecules are always translated into the same amino acid sequences, independent of the source of the translation machinery. At the start, the answer was thought to be yes, for it was hard to imagine how the code could change during the course of evolution. By now the initial expectations that the genetic code for chromosomal DNA would prove to be universal have been rigorously confirmed in a large variety of organisms, ranging from the simplest prokaryotes to the most complex eukaryotes. An interesting exception, however, occurs in the genetic code used by the DNA from mitochondria. Although for many years it was believed that DNA is located only in the nucleus, by the early 1960s it had become clear that the cytoplasmic organelles, the *mitochondria* and the *chloroplasts*, both possess their own unique DNA molecules. Now it is generally believed that mitochondria and chloroplasts represent the descendents of primitive bacterial cells that became symbiotically engulfed by primitive ancestors of the present eukaryotic organisms and increased the host's ATP-generating capacity (Chapter 22). For the most part, the genetic code used by mitochondria is identical with that used by nuclear DNA. However, UGA, a stop codon for nuclear DNA, is read as tryptophan in mitochondria. In addition, mitochondrial AUA is read as methionine, whereas AUA is read as isoleucine in nuclear DNA.

These differences are due to the relatively small number of different tRNAs coded by mitochondrial DNA. Only 22 different tRNA species are present in mitochondria, in contrast to the more than 40 tRNAs that are available for ordinary translation. In many cases, just the first two bases in a codon are actually read, with the base in the third position playing no role in the tRNA selection process.

Average-Sized Genes Contain at Least 1200 Base Pairs

Because all codons were found to contain three base pairs, it was obvious that the number of base pairs in a gene must be at least three times the number of amino acids in its respective polypeptide. An average-sized protein of 400 amino acids was thus thought to require a section of DNA consisting of some 1200 nucleotide pairs. Because this number was found to

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicant:	**Periannan Senapathy**	**Group Art Unit:**	1655
Serial No.:	09/431,451	**Examiner:**	Sisson, B.
Filed:	November 1, 1999	**Attorney Docket No.:**	**34623.004**
For:	Method for Amplifying Sequences from Unknown DNA		

BOX: FEE
Assistant Commissioner for Patents
Washington, D.C. 20231

RULE 132 DECLARATION OF DR. PERIANNAN SENAPATHY

I, Periannan Senapathy, do hereby declare and state as follows:

1. I am the sole named inventor of the above-identified patent application and am the sole inventor of the invention described and claimed therein. As the sole inventor, I am intimately familiar with the application and its contents.

2. I have read the Office Action dated February 1, 2000 in the above-identified patent application. The purpose of this Declaration is to show that the specification of the application as filed fully describes the invention as claimed. All of the experiments described in this Declaration were either performed personally by me or performed at my direction and under my personal supervision. The following experiments demonstrate that the above-identified patent application fully enables the invention as broadly as it is claimed.

3. Usually a PCR amplification of a DNA sequence between two fixed primers is carried out at an annealing temperature (Tm) about 5°C lower than the lower Tm of the two primers, where the Tm of the two primers are close to each other. The temperature of melting or annealing referred to in the patent specification (at page 17, lines 1-20) is the standard range of temperature of melting/annealing in a standard PCR protocol, which is about 5 degrees less than the predicted Tm average of the first primer and the second primer up to about 5-10 degrees higher than the Tm average. Within this range, the temperature can

be adjusted (reduced or increased) to fine tune the specificity of the reaction in particular instances. The higher temperatures within this range are what is generally referred to as "stringent" temperature conditions. Non-specific amplification can be avoided by increasing the temperature of melting, thereby making the reaction more "stringent" and hence, more "specific." Adjusting the annealing temperature of a PCR reaction to yield more "stringent" conditions is well known to those skilled in the art and is done routinely. In short, performing the PCR on any given sample is largely an empirical science; the various parameters are adjusted within generally known ranges to yield the cleanest amplification possible. This empirical approach to PCR is a necessary and inescapable aspect of the PCR itself. In the PCR experiments described below, I increased the temperature within the standard range of the PCR reaction, *i.e.*, to stringent conditions, so that non-specific amplification of DNA can be avoided.

4. A systematic approach was taken for demonstrating that the DNA amplification and sequencing approach using partly-fixed primers as recited in the claims (Claims 1-25 being directed specifically to sequencing; Claims 26-30 being directed to amplification) works as described in the subject patent application, and that it can be directly applied to the amplification and sequencing of long DNA molecules.

5. The partly-fixed primers of the present invention consist of oligonucleotide primers in which many positions have completely random nucleotides. This may require an increased concentration of primer in the PCR amplification reaction as compared to a conventional PCR reaction. Therefore, one of the main aims of the experiment described in the paragraphs which follow was to analyze the primer concentration that is needed for amplifying a target region within the template DNA.

6. First, a plasmid DNA (pGEM, Promega Corporation, Madison, Wis.), with a length of approx. 6000 nucleotides, was used as the template DNA. A sequence region from the plasmid DNA was chosen and first and second fixed primers were designed as follows:

first primer (fixed):	5'–AGA AGC AGC CGA AGA AAC–3'
second primer (fixed):	5'–GGC ACC CCC GAC AC–3'

A plurality of partly-fixed second primers was then designed in such a manner that each primer in the plurality contained approximately 50% random nucleotides. The sequence of each primer in the plurality was 5'–NNN NNN NNC GAC AC–3'; *i.e.*, each individual primer contained fixed nucleotides at 6 positions and random nucleotides at 8 positions. The predicted Tm average for the first fixed primer and the second fixed primer was 57.5° C (59° + 56° ÷ 2). A PCR reaction was carried out at 60° C Tm using pGEM DNA as the template and various combinations of the above-described primers. Various first, fixed primer concentrations and partly-fixed second primer concentrations were used in separate amplification reactions, and the PCR amplified products were analyzed by electrophoresis on a 2.0% agarose gel.

7. The results (Figure 1, attached and incorporated by reference) show that even a 1 X concentration of a partly-fixed second primer, compared to the regular primer concentration normally used in the standard PCR protocol, works well. The ideal concentration was found to be 1 X for the first primer and 10 X for the second primer. Only one band corresponding to the positive control band was seen in the experiments. This experiment shows that when using a plurality of primers, each primer having a region of fixed sequence and a region of randomized sequence (as recited in the claims), the concentration parameters can be adjusted (within reasonable and often-encountered limits) to yield clean amplification, and that a normal Tm range can be used.

8. The next experiment done was to examine the following two parameters: 1) varying the number of fixed nucleotides within a partly-fixed second primer and gauging what effect that yields; and 2) using a significantly longer DNA as the template. The aim is to show that variations of these two parameters will still enable the amplification of the target DNA by the claimed method.

9. Therefore, the next experiment was conducted with a significantly longer template DNA, namely *E. coli* DNA, which is approximately 5 million base pairs long. The complete sequence of the *E. coli* genome has been recently elucidated. Three separate regions from this DNA were chosen for PCR amplification by the presently claimed method. These regions are sequences bounded by the following nucleotide locations: Region 1: 15738 and 16281; Region 2: 7789 and 8288; Region 3: 15222 and 15673.

10. For each region, a set of 5 primers were made: 1) A first, fixed primer; 2) A second, fixed primer; 3) a plurality of partly-fixed second primers with 8 nucleotides fixed and 8 nucleotides fully randomized; 4) a plurality of partly-fixed second primers with 6 nucleotides fixed and 10 nucleotides fully randomized; and 5) a plurality of partly-fixed second primers with 5 nucleotides fixed and 11 nucleotides fully randomized. The goal here is to show that varying the number of fixed and random nucleotides within the plurality of partly-fixed primers as stated in the patent specification will still allow the invention to function as described in the body of the patent.

11. Another parameter that was tested was the primer concentration. The following parameters were tested in PCR amplification reactions: a standard template DNA concentration, a 50X first primer concentration, and 100X and 1000X second primer plurality concentration.

12. The primers used for the three regions are listed below:

<u>Region 1 (Average Temperature of melting/reannealing is 65.5° C)</u>:

First primer:
5'–ACG CGC CGC AGT TAC G–3'

Second primer (positive control):
5'–ATG GCC CGC TGC TTC C–3'

Partly-fixed second primer 1 (5 fixed nucleotides and 11 random nucleotides):
5'–NNN NNN NNN NGC TTC C–3'

Partly-fixed second primer 2 (6 fixed nucleotides and 10 random nucleotides):
5'–NNN NNN NNN NGC TTC C–3'

Partly-fixed second primer 3 (8 fixed nucleotides and 8 random nucleotides):
5'–NNN NNN NNC TGC TTC C–3'

Region 2 (Average Temperature of melting/reannealing is 62.5° C):

First primer:
5'–ACC GCC TGG CTG TGG A–3'

Second primer (positive control):
5'–CGC GAT GTC CCC AGT G–3'

Partly-fixed second primer 1 (5 fixed nucleotides and 11 random nucleotides):
5'–NNN NNN NNN NNC AGT G–3'

Partly-fixed second primer 2 (6 fixed nucleotides and 10 random nucleotides):
5'–NNN NNN NNN NCC AGT G–3'

Partly-fixed second primer 3 (8 fixed nucleotides and 8 random nucleotides):
5'–NNN NNN NNC CCC AGT G–3'

Region 3 (Average Temperature of melting/reannealing is 61° C):

First primer:
5'–CCG GCG AGC ACA ACA G–3'

Second primer (positive control):
5'–GCA GCC GCT TCA GGA G–3'

Partly-fixed second primer 1 (5 fixed nucleotides and 11 random nucleotides):
5'–NNN NNN NNN NNA GGA G–3'

Partly-fixed second primer 2 (6 fixed nucleotides and 10 random nucleotides):
5'–NNN NNN NNN NCA GGA G–3'

Partly-fixed second primer 3 (8 fixed nucleotides and 8 random nucleotides):
5'–NNN NNN NNT TCA CGA G–3'

13. The aim here was to carry out a series of experiments, where the temperature of reannealing was raised gradually within the normal range used in the PCR protocol. The first experiment was conducted at an annealing temperature (Tm) within the normal range for all the three sets of experiments, namely 60° C. This temperature, however, is fairly stringent for the experiment with Region 3 (average Tm = 61° C), and not so stringent for Regions 1 and 2 (average Tm = 65.5° C and 62.2° C, respectively). The DNA from each of the three regions were PCR amplified in separate experiments with whole *E. coli* DNA as template. For each region, the PCR amplification between the first fixed primer and the second fixed primer was tested as a positive control. The PCR amplification between the first primer and the three pluralities of partly-fixed second primers were tested in separate PCR reactions. Standard PCR protocols were followed, and the amplified products were analyzed by electrophoresis on a 2% agarose gel.

14. The results (Figure 2, attached and incorporated by reference) indicate that the expected band corresponding to the positive control was present in almost all the experiments in each of the three DNA regions, and that there is no significant background. There were only a few bands other than the expected band. These bands are predicted and expected to be formed, as stated in the specification, from the priming between the juxtaposed second primers. The idea, as stated in the patent specification, is that even if bands other than the expected band occur in the PCR products, sequencing would be done using the first primer which is specific only to the particular DNA band amplified from the first primer location in the target DNA. Thus, the sequencing reaction should produce only the sequence of the Specific expected DNA fragment.

15. Varying the concentration of the second plurality of primers (100X and 1000X) does not produce significant variation in the PCR products. It indicates that even lower concentrations of the second primers will work with equal success.

16. The PCR products from Region 3 were chosen for DNA sequencing because the experiment from this region happens to have the closest predicted Tm to that of the actual PCR amplification reaction, which is fairly stringent for this particular region. The PCR products were directly sequenced using the first primer from Region 3 as the sequencing primer.

17. As shown in the results from the experiment with the Region 3 (Figure 3, attached and incorporated by reference), the sequence obtained with the plurality of partly-fixed second primers nos. 1, 2, & 3, are the same as that generated by the positive control (the amplification product between the first fixed primer and the second fixed second primer). In this case, the Tm used (60° C) in the experiment is only slightly lower than that predicted for the fixed primers (61° C), and thus the amplification between the first

primer and the partly-fixed primers has been perfectly specific for the target region in the template DNA. Figure 4, attached and incorporated by reference, shows a comparison of these four sequences (using a computer program) to that of the whole *E. coli* sequence, indicating that the sequences are from the chosen Region 3.

18. The results show that the presence of a few other DNA fragments (bands) in the PCR products does not affect sequencing of the specific fragment amplified from the target region.

19. The experiments to this point show clearly that the specific DNA amplification and sequencing of the target DNA works very well with a plurality of partly-fixed primers. The success of the partly-fixed primers with 5 fixed and 11 random nucleotide primers and with 8 fixed and 8 random nucleotide primers indicates that the use of 4 nucleotide and 10 nucleotide fixed second primer with the rest random sequence will also work as well. Furthermore, these experiments indicate that changing the position of the fixed nucleotides and random nucleotides within the primers will not adversely affect the process.

20. Next, the effect of concentration of the plurality of partially fixed, partially randomized primers on the ability to sequence the resultant amplification products was explored. The number of individual primers within the plurality of primers that will bind to the template with complete, full-length complementarity depends upon the number of randomized bases. The method will function succesfully with one or two base pair mismatches within the randomized portion of the primer, especially if they are at the 5' end, because the primer will still be able to bind with full complementarity over a 13-14 base pair length at the 3' end of the individual primer. Thus, a lesser amount of the primer than needed to compensate mole to mole may be sufficient to achieve a suitable quantity of amplified target for sequencing.

21. Thus, the following experiment was performed to show that increasing the concentration of the randomized primers yields a concomitant increase in the quantity of the target amplified DNA fragment. Here, an experiment was conducted wherein the PCR amplification was carried out with and without increasing the concentration of the randomized second primer with a DNA template of 12.5 KB (a plasmid clone containing a human gene insert, kindly provided by Prof. M. Mahadevan of the University of Wisconsin, Department of Medical Genetics). Figure 6A, attached and incorporated by reference, shows that the full-length, 16-mer primer (lane 1), and the 16-mer primer with 2 Ns at the 5' end instead of the regular sequence (lane 2), and the 16-mer primer with 4 Ns at the 5' end (lane 3), were able to amplify the target DNA fragments at the standard concentrations of the primer, while the 16-mer primers with 6 Ns and 8 Ns (lane 4 & 5) did not produce a visible band. This means that the primer concentrations conventionally used in PCR represents a very large excess of the primer (4^4 or ~256-fold).

22. Next, the concentration of the primer plurality was increased to adjust for the reduced amount of individual primer species within the plurality having full complementarity to the target (and allowing for one to two mismatches within the randomized portion). Figure 6B, attached and incorporated by reference, shows that the 16-mer primer with 6 Ns at the 5' end (10-fold excess, lane 4) and the 16-mer primer with 8 Ns at the 5' end (120-fold excess, lane 5) also amplified the target DNA fragment. The 16-mer primer with 10 Ns at 250-fold excess (lane 7) did not amplify the product, which indicates that a higher concentration of the primer was needed. This result indicates that target-specific amplification can be obtained by increasing the concentration of the plurality of partially-fixed, partially-randomized primers to a reasonable extent (predicted by the number of mismatches). As illustrated above, in paragraphs 16-18, these specific amplification products are quite easily sequenced using conventional, automated methods. The results also show that the band intensity was reduced as the number of Ns was increased (see Figure 6B). However, even an 8-fixed/8-random base primer was able to generate the correct band with a reduced intensity. Furthermore, this experiment shows that the band intensity can be improved by increasing the quantity of the primer plurality used in the amplification reaction.

23. The experiments described above demonstrate the ability to amplify an unknown DNA downstream of a known sequence within a template DNA using randomized primers, and to sequence the amplified target fragments without gel purifying the fragment. However, it would be far more advantageous if a DNA template of completely unknown sequence could be cycle-sequenced directly using the randomized primers. The concept underlying the subject invention is to compute the length of the template DNA in which a given fixed portion of a randomized primer will bind once, on average. For example, a 5-base fixed primer will have one binding site, on average, in a template DNA of 1024 bases (approximately 1 KB). A 7-base fixed primer will have one binding site, on average, within a template DNA of about 16 KB. This means that a plurality of primers having a fully randomized portion and a 7-base fixed-sequence portion will contain a primer species with full-length complementarity and which will bind at a site where the 7-base fixed sequence portion binds on the template DNA. Statistically, there will be only one occurrence of the 7-base fixed sequence portion within a 16 KB long template DNA, and one of the primers within the plurality will bind only at this binding site within the 16 KB stretch of DNA. Thus, the claimed method provides a complementary, or nearly complementary, full-length primer species from the plurality of partly-randomized, partly-fixed primers which will bind specifically to the template DNA, and which can then be utilized directly for cycle-sequencing an approximately 500 base pair-long portion of the template DNA. The number of primer species with tolerable mismatches within the randomized primers is sufficient to enable the correct sequencing of the template DNA, as revealed by the results discussed below. See paragraph 25 for further description of the experiment illustrated in Figure 6B.

24. If the partly fixed randomized primers can be used to cycle sequence a short DNA template of known (as shown above for the pGEM DNA), then this basic principle can be extended and expanded to sequence longer DNA templates of unknown sequence (such as the *E. coli* DNA discussed above). This technique can then be used directly to cycle sequence large portions of genomic DNA by combining the direct sequencing technique with the direct amplification of unknown genomic DNA described above.

25. The hypothesis of paragraph 24 was shown to be true using a plasmid template DNA (pGEM) in the following experiment. The known sequence of the plasmid DNA was analyzed to choose a few cycle-sequencing primers. Full-length primers were made with these sequences. Each of the primers was also randomized at its 5' end to varying extents to yield various pluralities of primers having partially randomized portions within each primer. For example, the primer sequence TGT AAA ACG ACG GCC AGT was randomized at the 5' end to give 2, 4, 6, and 8 Ns, namely:

Full Length Primer: TGT AAA ACG ACG GCC AGT

GT2: 5' NNT AAA ACG ACG GCC AGT 3' (2 Degenerate bases on the 5 prime end)
Signal Strength: 2270

GT4: 5' NNN NAA ACG ACG GCC AGT 3' (4 Degenerate bases on the 5 prime end)
Signal Strength: 1834

GT6: 5' NNN NNN ACG ACG GCC AGT 3' (6 Degenerate bases on the 5 prime end)
Signal Strength: 1082

GT7: 5' NNN NNN NNG ACG GCC AGT 3' (8 Degenerate bases on the 5 prime end)
Signal Strength: 354

The sequencing results from these experiments are shown in Figures 5A, 5B, 5C, and 5D, respectively (attached and incorporated by reference). As can be seen from the printout from the sequencer, these reactions, performed according to the method recited in the claims, yields perfectly acceptable sequencing results. This experiment shows that specific and reproducible sequencing results can be obtained using a collection of primers wherein each primer within the collection has a fixed portion and a random portion.

26. Lastly, the strength of the sequencing signals from the experiment described in paragraph 25 were compared with the respective number of randomized bases (Ns) in the sequencing primers. As shown below, the signal strength is inversely proportional to the number of randomized bases:

GT2: 5' NNT AAA ACG ACG GCC AGT 3' 2 Ns - Signal Strength: 2270

GT4: 5' NNN NAA ACG ACG GCC AGT 3' 4 Ns - Signal Strength: 1834

GT6: 5' NNN NNN ACG ACG GCC AGT 3' 6 Ns - Signal Strength: 1082

GT7: 5' NNN NNN NNG ACG GCC AGT 3' 8 Ns - Signal Strength: 354

However, even the lowest signal strength generated a clearly readable sequence. Furthermore, by increasing the concentration of the primers having randomized base pairs, the signal strength was also increased in a predictable manner (data not shown).

27. The undersigned declares that all statements made herein of his own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment or both, under 5 1001, Title 18, of the U.S.C., and that such willful false statements may jeopardize the validity of the application or any patent issuing thereon.

Date: 27 July 2000

P Senapathy

Periannan Senapathy, Ph.D.

Legend for Figure 1:

Technique using a plasmid DNA: The template DNA used was from the plasmid pGEM. Standard conditions: primer 1 - 6.25 pm, primer 2 - 6.25 pm, template - 50 ng. Each reaction contains -1.5 units of Taq, 10 mM Tris-HCl, 50 mM KCl, 1.5 mM MgCl2, 200 mM of each dNTP and stabilizers, including BSA. The following PCR amplification protocol was used: 1 cycle at 95° C for 5 minutes, followed by 35 cycles of: 95° C for 2 minutes, 60° C for 1 minute; and 72° C for 2 minutes. Hold at 4° C. The enzyme used was Taq polymerase. Experiment 1: Vary 2nd primer and everything else is standard: Lane 1 - 1x 2nd primer, Lane 2 - 10x 2nd primer, Lane 3 - 100x 2nd primer, Lane 4 - 1000x 2nd primer. Experiment 2: Vary 1st primer and everything else is standard: Lane 5 - lx 1st primer, Lane 6 - 10x lst primer, Lane 7 - 100x lst primer, Lane 8 - 1000x 1st primer, Lane 9 - 50 base-pair ladder, Lane 10 - positive control between the 1st fixed primer and the 2nd fixed primer.

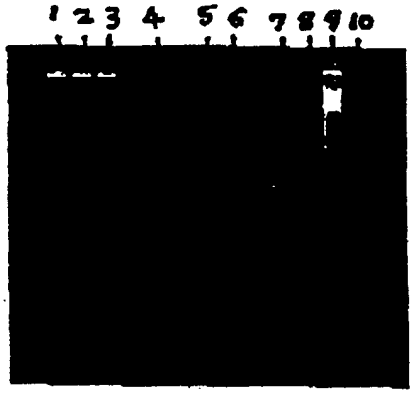


FIGURE 1

Legend for Figure 2:
Technique with complete *E. coli* DNA: Template DNA - complete *E. coli* genomic DNA. Standard conditions: primer 1 - 6.25 pm, primer 2 - 6.25 pm, template - 50 ng. The following PCR amplification protocol was used: 1 cycle at 95° C for 5 minutes, followed by 35 cycles of: 95° C for 2 minutes, 60° C for 1 minute; and 72° C for 2 minutes. Hold at 4° C. The enzyme used was Taq polymerase and the reaction conditions were as described for Fig. 1.

Region 1: Lanes 1-8:
Lane 1 - positive control between 1st fixed primer and 2nd fixed primer for region 1.
Lane 2 - negative control (everything else except the primers)
Lane 3 - partly fixed 2nd primer 1 (100x concentration)
Lane 4 - partly fixed 2nd primer 2 (100x concentration)
Lane 5 - partly fixed 2nd primer 3 (100x concentration)
Lane 6 - partly fixed 2nd primer 1 (1000x concentration)
Lane 7 - partly fixed 2nd primer 2 (1000x concentration)
Lane 8 - partly fixed 2nd primer 3 (1000x concentration)

Region 2: Lanes 9-15:
Lane 9 - positive control between 1st fixed primer and 2nd fixed primer for region 2.
Lane 10 - partly fixed 2nd primer 1 (100x concentration)
Lane 11 - partly fixed 2nd primer 2 (100x concentration)
Lane 12 - partly fixed 2nd primer 3 (100x concentration)
Lane 13 - partly fixed 2nd primer 1 (1000x concentration)
Lane 14 - partly fixed 2nd primer 2 (1000x concentration)
Lane 15 - partly fixed 2nd primer 3 (1000x concentration)

Region 3: Lanes 16-22:
Lane 16 - positive control between 1st fixed primer and 2nd fixed primer for region 3.
Lane 17 - partly fixed 2nd primer 1 (100x concentration)
Lane 18 - partly fixed 2nd primer 2 (100x concentration)
Lane 19 - partly fixed 2nd primer 3 (100x concentration)
Lane 20 - partly fixed 2nd primer 1 (1000x concentration)
Lane 21 - partly fixed 2nd primer 2 (1000x concentration)
Lane 22 - partly fixed 2nd primer 3 (1000x concentration)

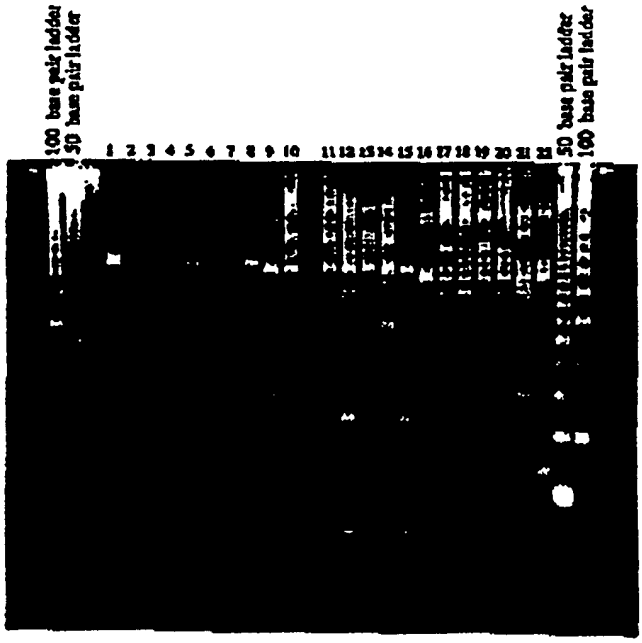

100 base pair ladder
50 base pair ladder
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
50 base pair ladder
100 base pair ladder

FIGURE 2

Legend for Figure 3:

DNA Sequencing of the PCR products. The PCR product from each amplification tube was filtered through the Pharmacia G-50 Sepharose column, and was directly sequenced using the ABI DNA PCR sequencing kit (enzyme and dye di-deoxy terminators). The PCR cycle protocol for sequencing was the following: 95° C for 3', 25 cycles of 95° C for 15 seconds, 50° C for 20 seconds, and 60° C for 4 minutes, 72° C for 10 minutes and hold at 4° C.

Sample 16: PCR product from the Lane 16 in Figure 2.
Sample 17: PCR product from the Lane 17 in Figure 2.
Sample 18: PCR product from the Lane 18 in Figure 2.
Sample 19: PCR product from the Lane 19 in Figure 2.

FIGURE 3 - SAMPLE 16

ABI PRISM™

Model 377
Version 3.0

16
Senapathy/Perry
16
Lane 18

Signal G:161 A:404 T:127 C:204
DT4%Ac{A Set-AnyPrimer}
NEW INSTRUMENT FILE!
Points 1284 to 8008 Base 1: 1284

Tue, May 6, 1997 6:31 PM
Tue, Apr 29, 1997 3:40 PM
Spacing: 9.68

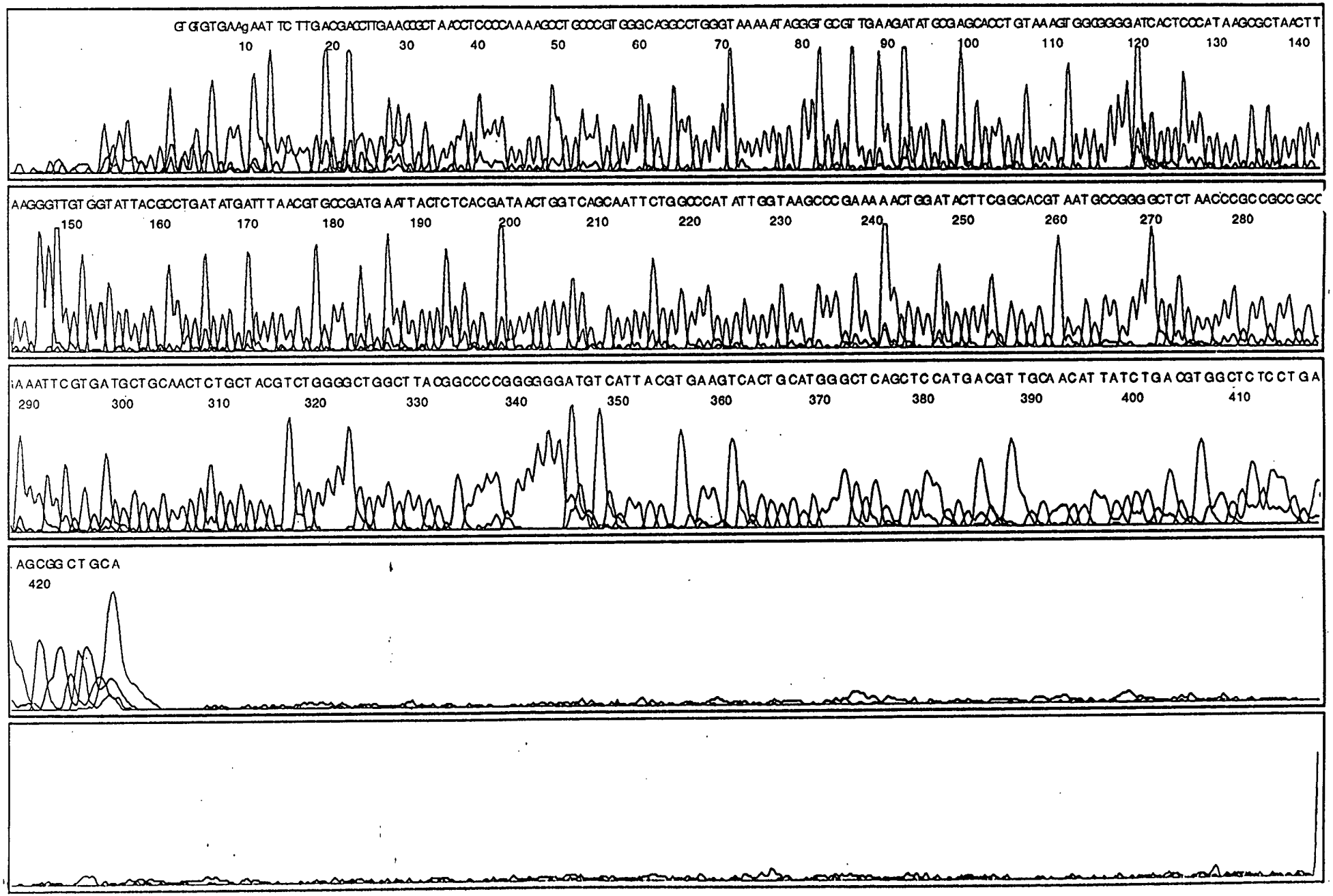


FIGURE 3 - SAMPLE 17

ABI PRISM™

Model 377
Version 3.0

17
Senapathy/Perry
17
Lane 20

Signal G:165 A:534 T:201 C:308
DT4%Ac{A Set-AnyPrimer}
NEW INSTRUMENT FILE!
Points 1284 to 8008 Base 1: 1284

Tue, May 6, 1997 6:31 PM
Tue, Apr 29, 1997 3:40 PM
Spacing: 9.69

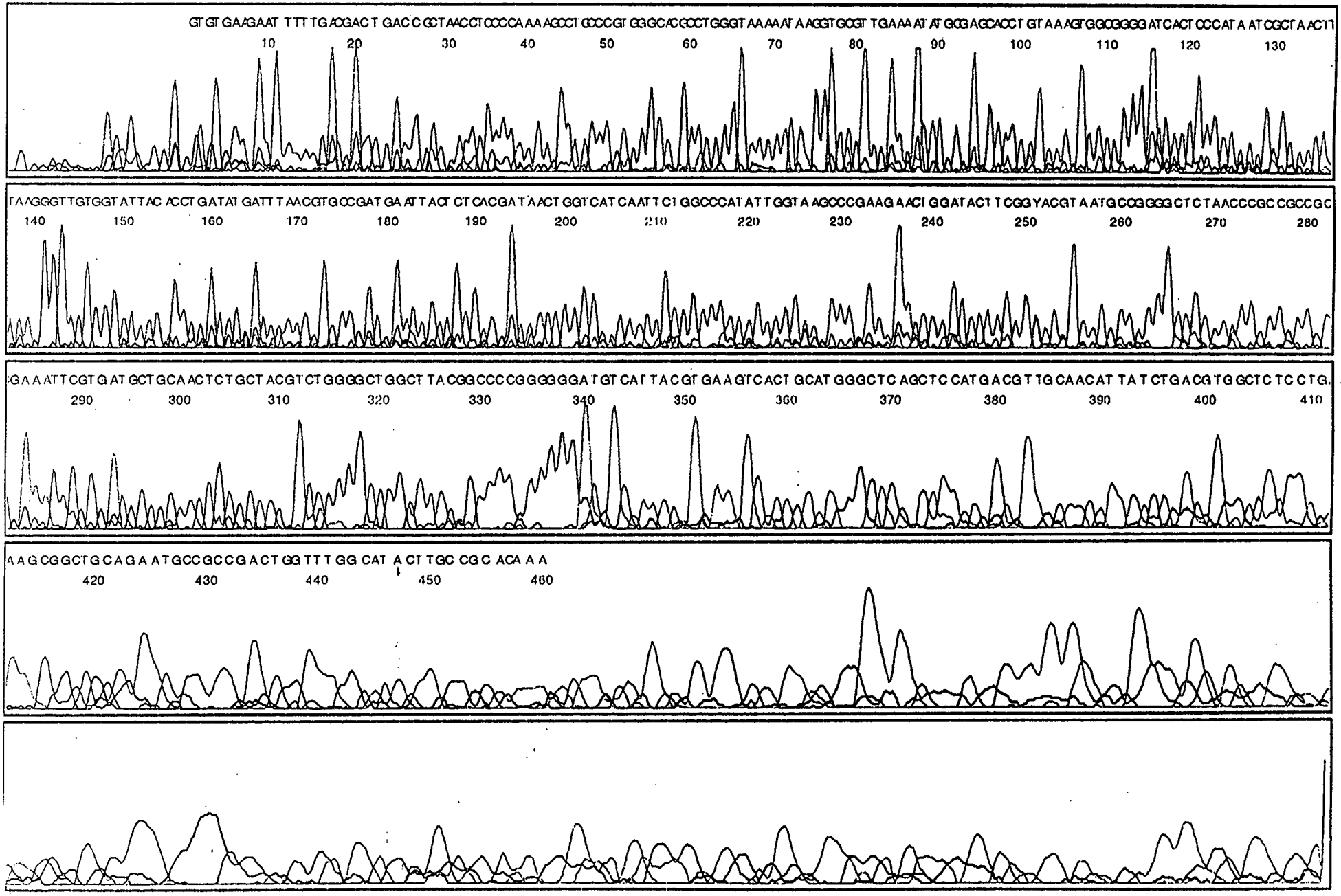


FIGURE 3 - SAMPLE 18

ABI PRISM™

Model 377
Version 3.0

18
Senapathy/Perry
18
Lane 21

Signal G:165 A:535 T:180 C:285
DT4%Ac{A Set-AnyPrimer}
NEW INSTRUMENT FILE!
Points 1251 to 8008 Base 1: 1251

Tue, May 6, 1997 6:31 PM
Tue, Apr 29, 1997 3:40 PM
Spacing: 9.60

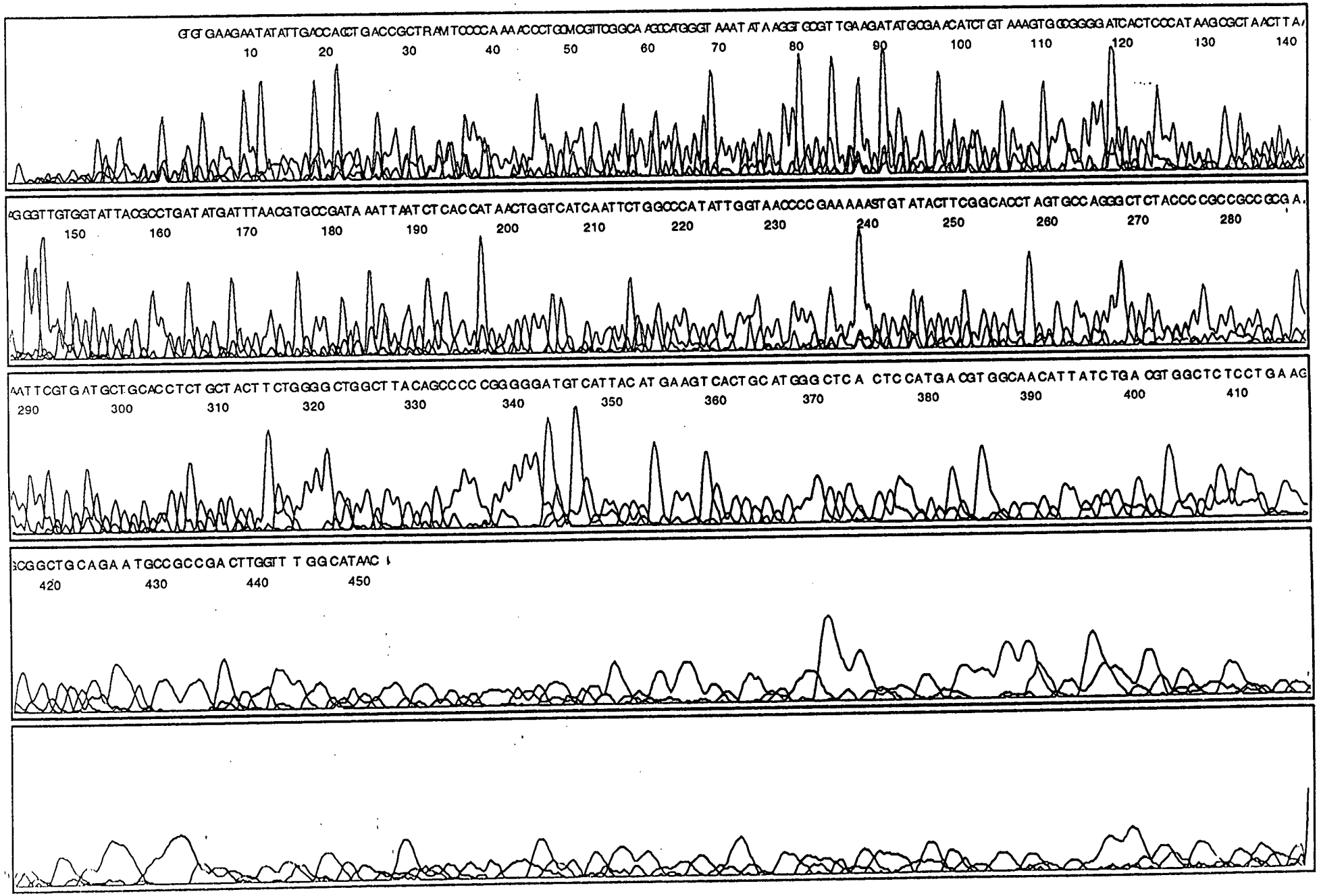


FIGURE 3 - SAMPLE 19

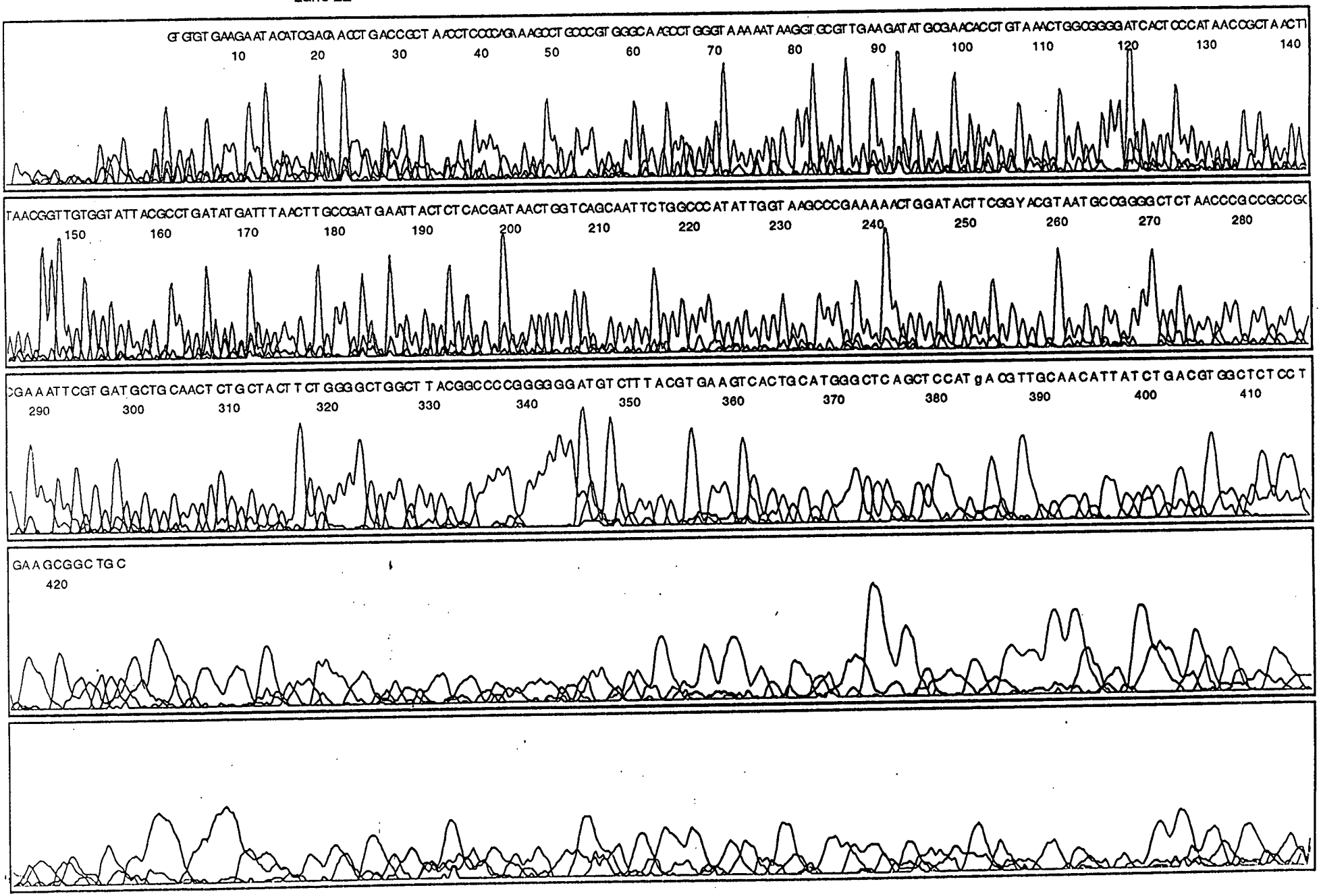
ABI PRISM™

Model 377
Version 3.0

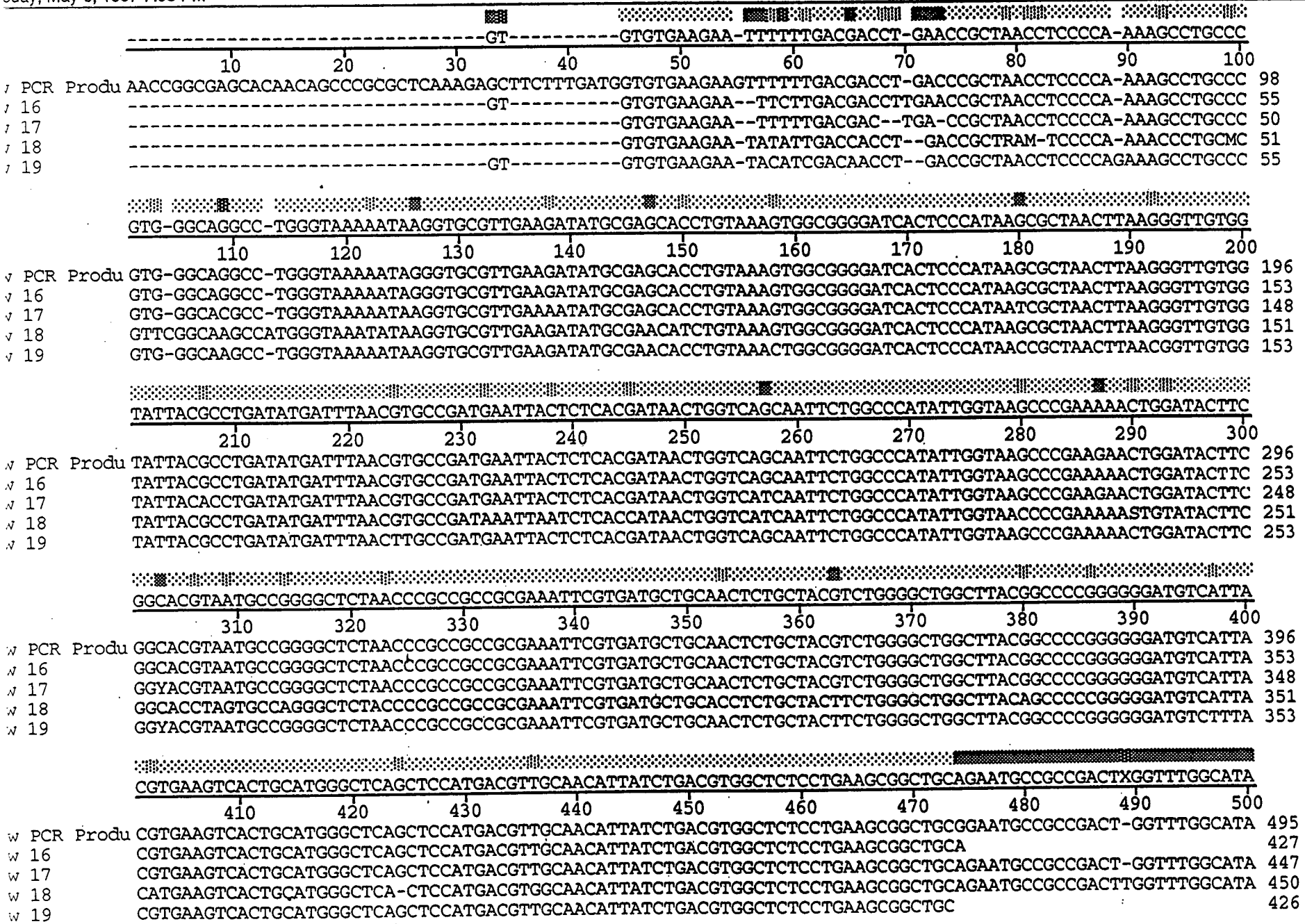
19
Senapathy/Perry
19
Lane 22

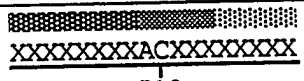
Signal G:90 A:295 T:100 C:154
DT4%Ac{A Set-AnyPrimer}
NEW INSTRUMENT FILE!
Points 1287 to 8008 Base 1: 1287

Tue, May 6, 1997 6:31 PM
Tue, Apr 29, 1997 3:40 PM
Spacing: 9.69



Legend for Figure 4:
Comparison of the sequences generated from PCR products (from Figure 3). The sequences were compared using a computer program for sequence comparison. Top row: *E. coli* genomic DNA sequence, New 16: Sample 16 from Figure 3 (positive control), New 17: Sample 17 from Figure 3 (partly-fixed 2nd primer with 5 fixed nucleotides), New 18: Sample 18 from Figure 3 (partly-fixed 2nd primer with 6 fixed nucleotides), New 19: Sample 19 from Figure 3 (partly-fixed 2nd primer with 8 fixed nucleotides).


----------GT----------GTGTGAAGAA-TTTTTTGACGACCT-GAACCGCTAACCTCCCCA-AAAGCCTGCCC
10 20 30 40 50 60 70 80 90 100
PCR Produ AACCGGCGAGCACAACAGCCCGCGCTCAAAGAGCTTCTTTGATGGTGTGAAGAAGTTTTTTGACGACCT-GACCCGCTAACCTCCCCA-AAAGCCTGCCC 98
16 ----------GT----------GTGTGAAGAA--TTCTTGACGACCTTGAACCGCTAACCTCCCCA-AAAGCCTGCCC 55
17 ----------GTGTGAAGAA--TTTTTGACGAC--TGA-CCGCTAACCTCCCCA-AAAGCCTGCCC 50
18 ----------GTGTGAAGAA-TATATTGACCACCT--GACCGCTRAM-TCCCCA-AAACCCTGCMC 51
19 ----------GT----------GTGTGAAGAA-TACATCGACAACCT--GACCGCTAACCTCCCCAGAAAGCCTGCCC 55
GTG-GGCAGGCC-TGGGTAAAAATAAGGTGCGTTGAAGATATGCGAGCACCTGTAAAGTGGCGGGGATCACTCCCATAAGCGCTAACTTAAGGGTTGTGG
110 120 130 140 150 160 170 180 190 200
PCR Produ GTG-GGCAGGCC-TGGGTAAAAATAGGGTGCGTTGAAGATATGCGAGCACCTGTAAAGTGGCGGGGATCACTCCCATAAGCGCTAACTTAAGGGTTGTGG 196
16 GTG-GGCAGGCC-TGGGTAAAAATAGGGTGCGTTGAAGATATGCGAGCACCTGTAAAGTGGCGGGGATCACTCCCATAAGCGCTAACTTAAGGGTTGTGG 153
17 GTG-GGCACGCC-TGGGTAAAAATAAGGTGCGTTGAAAATATGCGAGCACCTGTAAAGTGGCGGGGATCACTCCCATAATCGCTAACTTAAGGGTTGTGG 148
18 GTTCGGCAAGCCATGGGTAAATATAAGGTGCGTTGAAGATATGCGAACATCTGTAAAGTGGCGGGGATCACTCCCATAAGCGCTAACTTAAGGGTTGTGG 151
19 GTG-GGCAAGCC-TGGGTAAAAATAAGGTGCGTTGAAGATATGCGAACACCTGTAAACTGGCGGGGATCACTCCCATAACCGCTAACTTAACGGTTGTGG 153
TATTACGCCTGATATGATTTAACGTGCCGATGAATTACTCTCACGATAACTGGTCAGCAATTCTGGCCCATATTGGTAAGCCCGAAAAACTGGATACTTC
210 220 230 240 250 260 270 280 290 300
PCR Produ TATTACGCCTGATATGATTTAACGTGCCGATGAATTACTCTCACGATAACTGGTCAGCAATTCTGGCCCATATTGGTAAGCCCGAAGAACTGGATACTTC 296
16 TATTACGCCTGATATGATTTAACGTGCCGATGAATTACTCTCACGATAACTGGTCAGCAATTCTGGCCCATATTGGTAAGCCCGAAAAACTGGATACTTC 253
17 TATTACACCTGATATGATTTAACGTGCCGATGAATTACTCTCACGATAACTGGTCATCAATTCTGGCCCATATTGGTAAGCCCGAAGAACTGGATACTTC 248
18 TATTACGCCTGATATGATTTAACGTGCCGATAAATTAATCTCACCATAACTGGTCATCAATTCTGGCCCATATTGGTAACCCCGAAAAASTGTATACTTC 251
19 TATTACGCCTGATATGATTTAACTTGCCGATGAATTACTCTCACGATAACTGGTCAGCAATTCTGGCCCATATTGGTAAGCCCGAAAAACTGGATACTTC 253
GGCACGTAATGCCGGGGCTCTAACCCGCCGCCGCGAAATTCGTGATGCTGCAACTCTGCTACGTCTGGGGCTGGCTTACGGCCCCGGGGGGGATGTCATTA
310 320 330 340 350 360 370 380 390 400
PCR Produ GGCACGTAATGCCGGGGCTCTAACCCGCCGCCGCGAAATTCGTGATGCTGCAACTCTGCTACGTCTGGGGCTGGCTTACGGCCCCGGGGGGGATGTCATTA 396
16 GGCACGTAATGCCGGGGCTCTAACCCGCCGCCGCGAAATTCGTGATGCTGCAACTCTGCTACGTCTGGGGCTGGCTTACGGCCCCGGGGGGGATGTCATTA 353
17 GGYACGTAATGCCGGGGCTCTAACCCGCCGCCGCGAAATTCGTGATGCTGCAACTCTGCTACGTCTGGGGCTGGCTTACGGCCCCGGGGGGGATGTCATTA 348
18 GGCACCTAGTGCCAGGGCTCTACCCCGCCGCCGCGAAATTCGTGATGCTGCACCTCTGCTACTTCTGGGGCTGGCTTACAGCCCCCGGGGGGATGTCATTA 351
19 GGYACGTAATGCCGGGGCTCTAACCCGCCGCCGCGAAATTCGTGATGCTGCAACTCTGCTACTTCTGGGGCTGGCTTACGGCCCCGGGGGGATGTCTTTA 353
CGTGAAGTCACTGCATGGGCTCAGCTCCATGACGTTGCAACATTATCTGACGTGGCTCTCCTGAAGCGGCTGCAGAATGCCGCCGACTXGGTTTGGCATA
410 420 430 440 450 460 470 480 490 500
PCR Produ CGTGAAGTCACTGCATGGGCTCAGCTCCATGACGTTGCAACATTATCTGACGTGGCTCTCCTGAAGCGGCTGCGGAATGCCGCCGACT-GGTTTGGCATA 495
16 CGTGAAGTCACTGCATGGGCTCAGCTCCATGACGTTGCAACATTATCTGACGTGGCTCTCCTGAAGCGGCTGCA 427
17 CGTGAAGTCACTGCATGGGCTCAGCTCCATGACGTTGCAACATTATCTGACGTGGCTCTCCTGAAGCGGCTGCAGAATGCCGCCGACT-GGTTTGGCATA 447
18 CATGAAGTCACTGCATGGGCTCA-CTCCATGACGTGGCAACATTATCTGACGTGGCTCTCCTGAAGCGGCTGCAGAATGCCGCCGACTTGGTTTGGCATA 450
19 CGTGAAGTCACTGCATGGGCTCAGCTCCATGACGTTGCAACATTATCTGACGTGGCTCTCCTGAAGCGGCTGC 426



```
                XXXXXXXXXACXXXXXXXX
                        510
PCR Produ CTTGCCGCACAAACACTT                                    513
16                                                              427
17        CTTGCCGCACAAA                                         460
18        ---------AC                                           452
19                                                              426
```

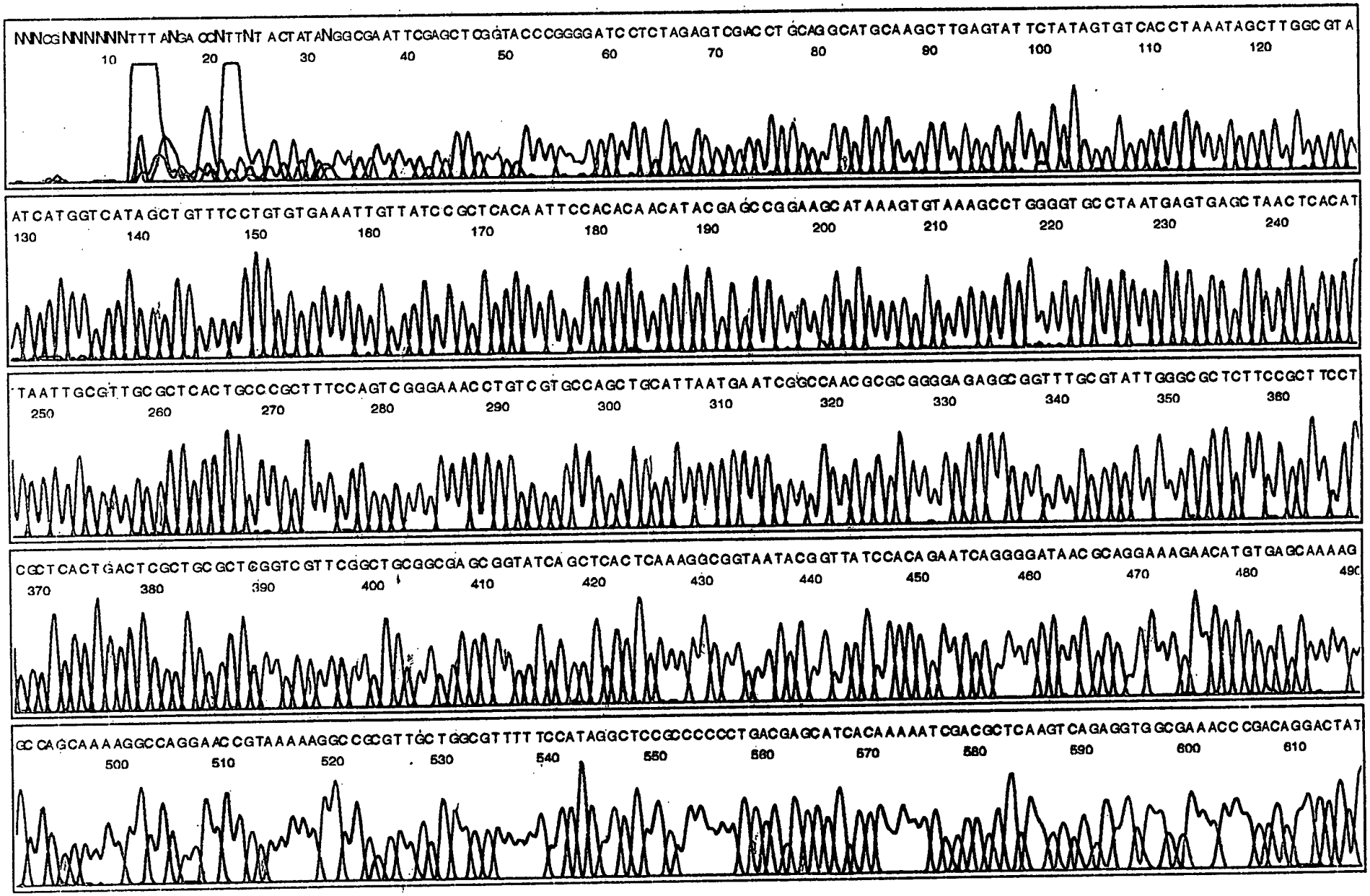
ABI PRISM™

Model 377
Version 3.2

GT2
Senapathy/Cris
GT2
Lane 39

Signal G:724 A:465 T:503 C:578
DT {BD Set Any-Primer}
dRhod
Points 979 to 9444 Base 1: 979

Tue, Aug 24, 1999 9:59 AM
Mon, Aug 23, 1999 1:51 PM
Spacing: 9.64



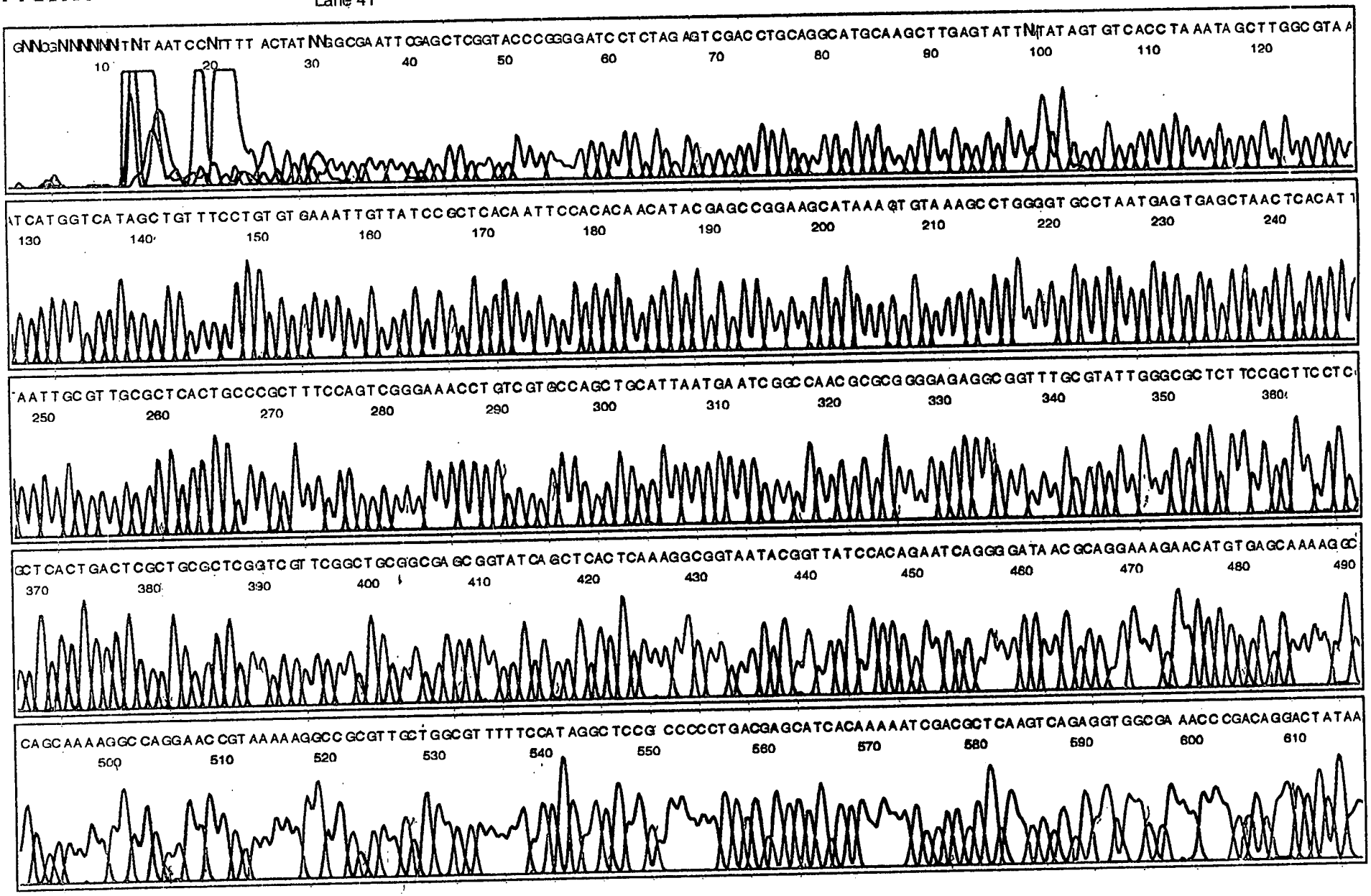
ABI PRISM™

Model 377
Version 3.2

GT4
Senapathy/Cris
GT4
Lane 41

Signal G:566 A:374 T:436 C:458
DT {BD Set Any-Primer}
dRhod
Points 983 to 9444 Base 1: 983

Tue, Aug 24, 1999 9:59 AM
Mon, Aug 23, 1999 1:51 PM
Spacing: 9.67



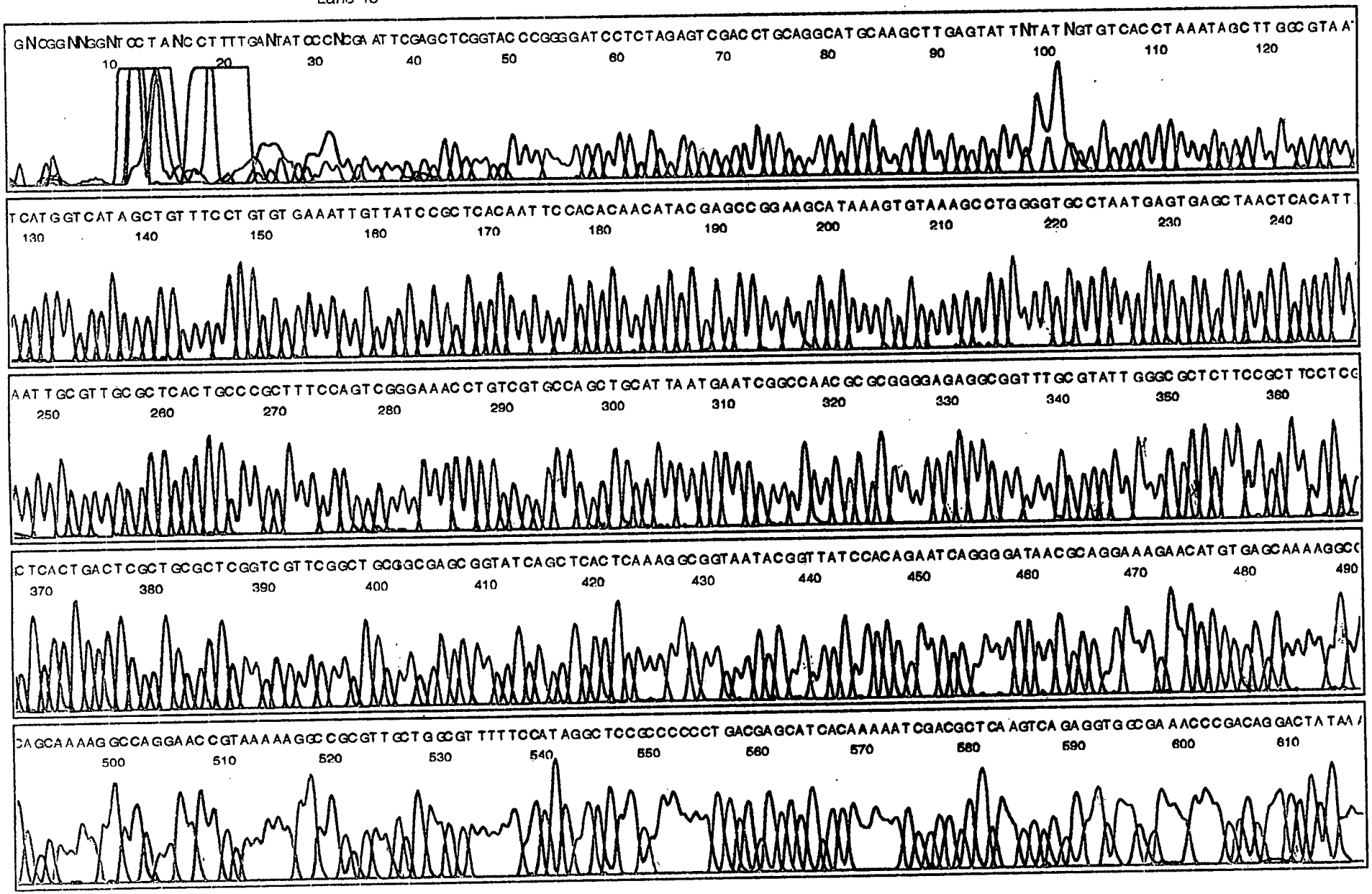
ABI PRISM™

Model 377
Version 3.2

GT6
Senapathy/Cris
GT6
Lane 43

Signal G:304 A:209 T:311 C:258
DT {BD Set Any-Primer}
dRhod
Points 988 to 9444 Base 1: 988

Tue, Aug 24, 1999 9:59 AM
Mon, Aug 23, 1999 1:51 PM
Spacing: 9.67



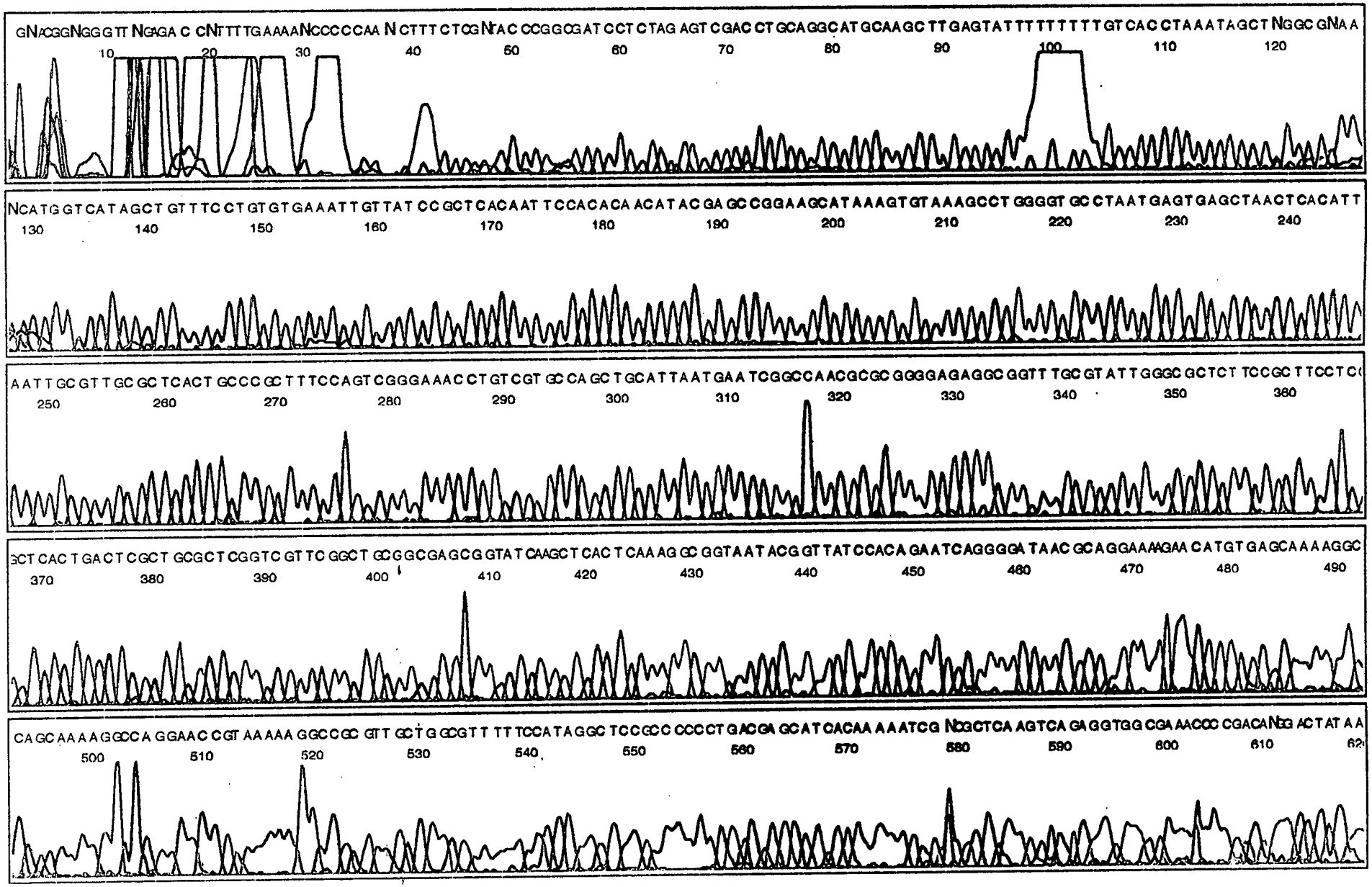
ABI PRISM™

Model 377
Version 3.2

GT7
Senapathy/Cris
GT7
Lane 44

Signal G:75 A:56 T:157 C:66
DT {BD Set Any-Primer}
dRhod
Points 990 to 9444 Base 1: 990

Tue, Aug 24, 1999 9:59 AM
Mon, Aug 23, 1999 1:51 PM
Spacing: 9.66



1 2 3 4 5

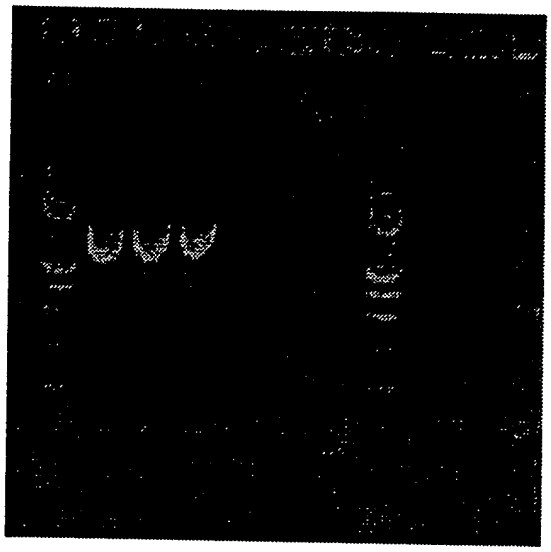


FIGURE 6 A

1 2 3 4 5 6 7

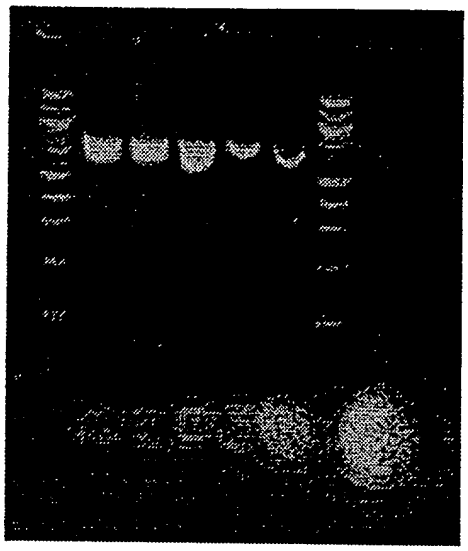


FIGURE 6 B

US005994058A

United States Patent [19]

Senapathy

[11] Patent Number: 5,994,058

[45] Date of Patent: Nov. 30, 1999

[54] **METHOD FOR CONTIGUOUS GENOME SEQUENCING**

[75] Inventor: **Periannan Senapathy**, Wheaton, Ill.

[73] Assignee: **Genome International Corporation**, Madison, Wis.

[21] Appl. No.: **08/406,545**

[22] Filed: **Mar. 20, 1995**

[51] **Int. Cl.**[6] **C12Q 1/68**; C07H 21/02; C12P 19/34; C12N 15/00

[52] **U.S. Cl.** **435/6**; 536/23.1; 536/24.3; 935/76; 935/77; 935/78

[58] **Field of Search** 435/6, 91.2; 536/22.1, 536/23.1, 24.3; 935/76, 77, 78

[56] **References Cited**

U.S. PATENT DOCUMENTS

5,104,792	4/1992	Silver et al.	435/6
5,106,727	4/1992	Hartley et al.	435/6
5,108,892	4/1992	Burke et al.	435/6
5,407,799	4/1995	Studier	435/6
5,487,985	1/1996	McClelland et al.	435/91.2
5,487,993	1/1996	Hermstadt et al.	435/172.3
5,508,169	4/1996	Deugau	435/6
5,731,171	3/1998	Bohlander	435/91.2

OTHER PUBLICATIONS

DNA Sequence—Journal DNA Sequencing and Mapping, vol. 2, issued 1992, P. Verhasselt et al., *DNA sequencing by a subcloning–walking strategy using a specific and semi–random primer in the polymerase chain reaction*, pp. 281–287.

Sarkar et al., DNA and Cell Biology 12 (7): 611–615 (1993).

Loh et al., Science 243: 217–220 (1988).

Volinia et al., CABIOS 5: 33–40 (1989).

Liu et al., Genomics 25: 674–681 (1995).

Kotler et al., PNAS 90: 4241–4245 (1993).

Caetano–Anollés et al., Mol. Gen. Genet. 235:157–165 (1992).

Nevinsky et al., Biochemistry 29:1200–1207 (1990).

Studier, F W., PNAS 86: 6971–6921 (1989).

Parker et al. Nucleic Acids Research 19(11):3055–3060.

Primary Examiner—W. Gary Jones

Assistant Examiner—Ethan Whisenant

[57] **ABSTRACT**

A new contiguous genome sequencing method is described which allows the contiguous sequencing of a very long DNA without need to be subcloned. It uses the basic PCR technique but circumvents the usual need of this technique for the knowledge two primers for contiguous sequencing, enabling the knowledge of only one primer sufficient. The present invention makes it possible to PCR amplify a DNA adjacent to a known sequence with which one primer can be made without the knowledge of the second primer binding site present in the unknown sequence. The present invention could thus be used to contiguously sequence a very long DNA such as that contained in a YAC clone or a cosmid clone, without the need for subcloning smaller fragments, using the standard PCR technique. It can also be used to sequence a whole chromosome or genome without any need to subclone it.

6 Claims, 5 Drawing Sheets

DeWitt Ross Docket:

34623.009

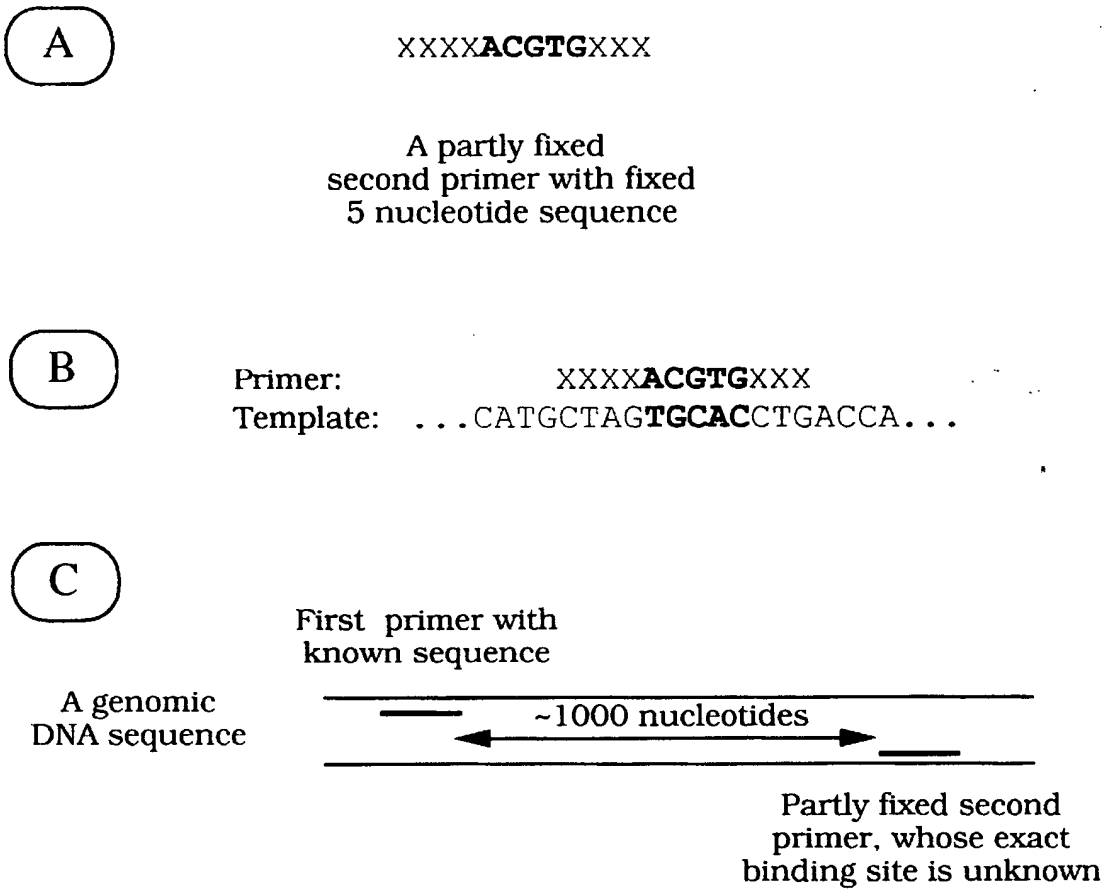


FIG. 1.

A) A partly random primer with a fixed sequence that is used as the second primer in the present invention. The length and composition of the fixed sequence and/or the random sequence (X's) and their positions within the primer, and the length of the primer itself can be adjusted to improve binding specificity and affinity, within the reaction conditions generally used in PCR protocols and DNA sequencing reactions.

B) The primer can bind to the template only at the sequence complementary to the fixed sequence in the primer.

C) This partly fixed second primer has a high specificity of binding at an ideal distance from the first known primer for PCR amplification and DNA sequencing.

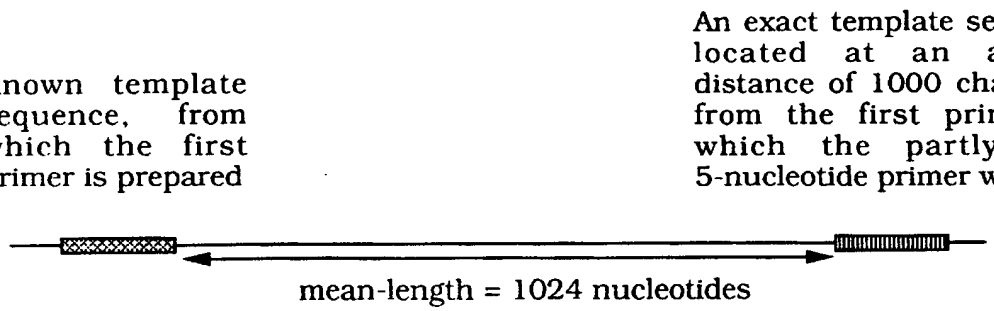


FIG.2. Precise binding of the partly fixed second primer to a specific sequence, located at an unknown distance from the first primer, but within a specified range of distance which is highly suitable for DNA amplification by the PCR technique and DNA sequencing by the standard techniques.

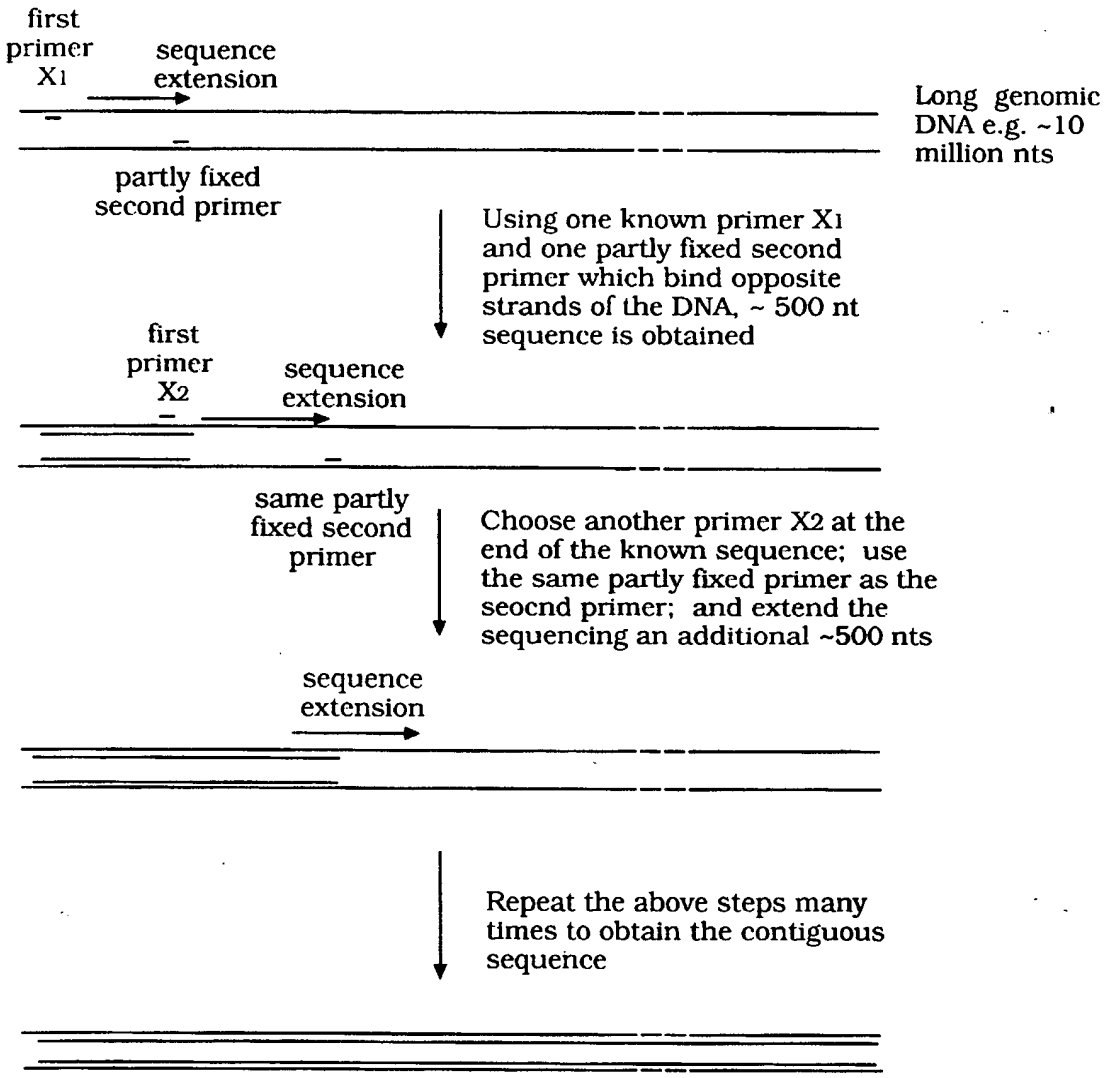


FIG. 3. Congituous genome-sequencing using a partly fixed primer as the second primer.

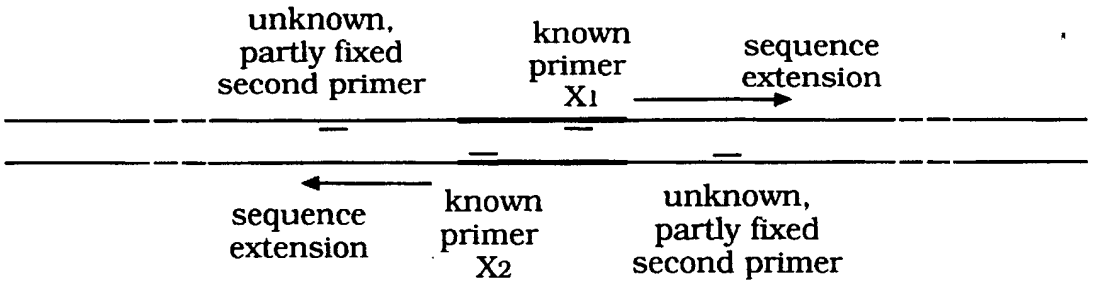


FIG. 4. Extension of DNA sequencing in both directions from a known sequence region (bold lines) in a very long DNA, using the same set of partly fixed second primers.

A

ATCATGTAAGTAGGC

An arbitrary but exact primer sequence of 15 nucleotides length, which would probabilistically have approximately one binding site in a genome of about one billion nucleotides.

B

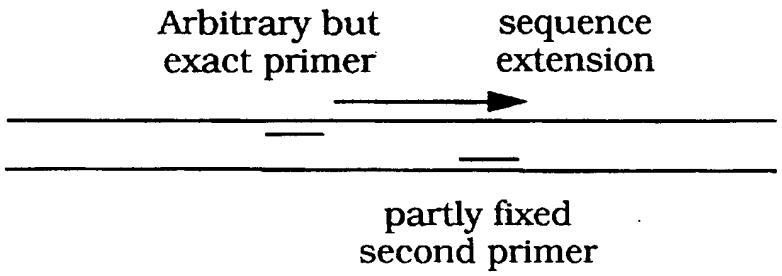


FIG. 5. Determining a DNA sequence from a long DNA of completely unknown sequence using an arbitrary but exact sequence primer and an unknown, partly fixed second primer.

METHOD FOR CONTIGUOUS GENOME SEQUENCING

BACKGROUND OF THE INVENTION

The present invention provides a method for contiguous sequencing of very long DNA using a modification of the standard PCR technique without the need for breaking down and subcloning the long DNA.

The PCR technique enables the amplification of DNA which lies between two regions of known sequence (K. B. Mullis et al., U.S. Pat. Nos. 4,683,202; 7/1987; 435/91; and 4,683,195, 7/1987; 435/6). Oligonucleotides complementary to these known sequences at both ends serve as "primers" in the PCR procedure. Double stranded target DNA is first melted to separate the DNA strands, and then oligonucleotide (oligo) primers complementary to the ends of the segment which is desired to be amplified are annealed to the template DNA. The oligos serve as primers for the synthesis of new complementary DNA strands, using a DNA polymerase enzyme and a process known as primer extension. The orientation of the primers with respect to one another is such that the 5' to 3' extension product from each primer contains, when extended far enough, the sequence which is complementary to the other oligo. Thus, each newly synthesized DNA strand becomes a template for synthesis of another DNA strand beginning with the other oligo as primer. Repeated cycles of melting, annealing of oligo primers, and primer extension lead to a (near) doubling, with each cycle, of DNA strands containing the sequence of the template beginning with the sequence of one oligo and ending with the sequence of the other oligo.

The key requirement for this exponential increase of template DNA is the two oligo primers complementary to the ends of the sequence desired to be amplified, and oriented such that their 3' extension products proceed toward each other. If the sequence at both ends of the segment to be amplified is not known, complementary oligos cannot be made and standard PCR cannot be performed. The object of the present invention is to overcome the need for sequence information at both ends of the segment to be amplified, i.e. to provide a method which allows PCR to be performed when sequence is known for only a single region, and to provide a method for the contiguous sequencing of a very long DNA without the need for subcloning of the DNA.

DNA sequencing is a technique by which the four DNA nucleotides (characters) in a linear DNA sequence is ordered by chemical and biochemical means. There are two techniques: 1) the chemical method of Maxam and Gilbert (A. M. Maxam, and W. Gilbert, "A new method of sequencing DNA." *Proceedings of the National Academy of Sciences, USA*, 74:560–564 (1977)), and the enzymatic method of Sanger and colleagues (F. Sanger, S. Nicklen, and A. R. Coulson, "DNA sequencing with chain-terminating inhibitors." 74:5463–5467 (1977)). In the chemical method, the DNA strand is isotropically labeled on one end, broken down into smaller fragments at sequence locations ending with a particular nucleotide (A, T, C, or G) by chemical means, and the fragments ordered based on this information. The four nucleotide-specific reaction products are resolved on a polyacrylamide gel, and the auto radiographic image of the gel is examined to infer the DNA sequence.

In the enzymatic method, the following basic steps are involved:

(i) annealing an oligonucleotide primer to a suitable single or denatured double stranded DNA template; (ii) extending the primer with DNA polymerase in four separate reactions, each containing one α-labeled dNTP or ddNTP (alternatively a labeled primer can be used), a mixture of unlabeled dNTPs, and one chain-terminating dideoxynucleoside-5'-triphosphate (ddNTP); (iii) resolving the four sets of reaction products on a high resolution polyacrylamide-urea gel; and (iv) producing an auto radiographic image of the gel that can be examined to infer the DNA sequence. Alternatively, fluorescently labeled primers or nucleotides can be used to identify the reaction products. Known dideoxy sequencing methods utilize a DNA polymerase such as the Klenow fragment of *E. coli* DNA polymerase, reverse transcriptase, a modified T7 DNA polymerase, or the Taq polymerase.

The PCR amplification procedure has been used to sequence the DNA being amplified (e.g. "Introduction to the AmpliTaq Cycle Sequencing Kit Protocol", a booklet from Perkin Elmer Cetus Corporation). The DNA could be first amplified and then it could be sequenced using the two conventional DNA sequencing techniques. Modified methods for sequencing PCR-amplified DNA have also been developed (e.g. Bevan et al., "Sequencing of PCR-Amplified DNA" *PCR Meth. App.* 4:222 (1992)). However, amplifying and sequencing using the PCR procedure requires that the sequences at the ends of the DNA (the two primer sequences) be known in advance. Thus, this procedure is limited in utility, and cannot be extended to contiguously sequence a long DNA strand. If the knowledge of only one primer is sufficient without anything known about the other primer, it would be greatly advantageous for sequencing very long DNA molecules using the PCR procedure. It would then be possible to use such a method for contiguously sequencing a long genomic DNA without the need for subcloning it into smaller fragments, and knowing only the very first, beginning primer in the whole long DNA.

In the currently existing methods for sequencing very long DNA of millions of nucleotides, the DNA is fragmented into smaller, overlapping fragments, and sub-cloned to produce numerous clones containing overlapping DNA sequences. These clones are sequenced randomly and the sequences assembled by "overlap sequence-matching" to produce the contiguous sequence. In this shot-gun sequencing method, approx. ten times more sequencing than the length of the DNA being sequenced is required to assemble the contiguous sequence. In the "directed" sequencing method, the linear order of the DNA clones has to be first determined by "physical mapping" of the clones.

There exists a contiguous DNA sequencing method called the "primer-walking" method using the Sanger's DNA polymerase enzymatic sequencing procedure. In this method, however, the DNA copying has to occur always from the template DNA during DNA sequencing. In contrast, in the PCR procedure, the target DNA amplified in the first rounds from the original input template DNA will function as the template DNA in subsequent cycles of amplification. After a certain cycles of amplification, the DNA sequencing reaction will be started by adding the sequencing "cocktail". Thus in the PCR reaction, only one copy of template DNA is theoretically sufficient to amplify into millions of copies, and therefore a very little genomic (or template) DNA is sufficient for sequencing. The advantage of DNA amplification that exists in PCR is lacking in the conventional Sanger procedure. Thus, this primer-walking method will require a larger amount of template DNA compared to the PCR sequencing method. Also, because the long DNA has a tendency to re-anneal back to duplex DNA, the sequencing gel pattern may not be as clean as in a PCR procedure, when a very long DNA is being sequenced. This may limit the

length of the DNA, that could be contiguously sequenced without breaking the DNA, using the primer-walking procedure. The PCR method also enables the reduction of non-specific binding of the primers to the template DNA because the enzymes used in these protocols function at high-temperatures, and thus allow "stringent" reaction conditions to be used to improve sequencing.

The present method of contiguous DNA sequencing using the basic PCR technique has thus many advantages over the primer walking method. Also, so far no method exists for contiguously sequencing a very long DNA using PCR technique. The present invention thus offers a unique and very advantageous procedure for contiguous DNA sequencing.

SUMMARY OF THE INVENTION

The present invention enables the amplification of a DNA stretch using the PCR procedure with the knowledge of only one primer. Using this basic method, the present invention describes a procedure by which a very long DNA of the order of millions of nucleotides can be sequenced contiguously, without the need for fragmenting and subcloning the DNA. In this method, the general PCR technique is used, but the knowledge of only one primer is sufficient, and the knowledge of the other primer is derived from the statistics of the distributions of oligonucleotide sequences of specified lengths.

Present DNA sequencing methods using the separation of DNA fragments on a gel has a limitation of resolving the products of length up to about 1000 nucleotides. Thus, in a single step, the sequence of a DNA fragment up to a length of only about 1000 nucleotides can be obtained by the two conventional DNA sequencing methods. A DNA sequence of a few nucleotides up to many thousand nucleotides can be amplified by the PCR procedure. Thus the PCR procedure can be combined with the DNA sequencing procedure successfully.

A primer is usually of length twelve nucleotides and longer. Let the sequence of one primer is known in a long DNA sequence from which the DNA sequence is to be worked out. From this primer sequence, a specific sequence of four nucleotides occurs statistically at an average distance of 256 nucleotides. It has been worked out by Senapathy that a particular sequence of four characters would occur anywhere from zero distance up to about 1500 characters with a 99.9% probability (P. Senapathy, "Distribution and repetition of sequence elements in eukaryotic DNA: New insights by computer aided statistical analysis," *Molecular Genetics (Life Sciences Advances)*, 7:53–65 (1988)). The mean distance for such an occurrence is 256 characters and the median is 180 characters. Similarly, a 5 nucleotide long specific sequence will occur at a mean distance of 1024 characters, with 99.99% of them occurring within 6000 characters from the first primer. The median distance for the occurrence of a 5-nucleotide specific sequence is ~730 nucleotides. Similarly, a particular 6 nucleotide long sequence will occur at a mean distance of 4096 nucleotides and a median distance of ~2800 nucleotides. A primer of known length, say length 14 can be prepared with a known sequence of 6 characters and the rest of the sequence being random in sequence. It means that any of the four nucleotides can occur at the "random" sequence locations. With a fixed 5, 6 or 7 nucleotide sequence within the second primer, a primer of length 12–18 can be prepared with high specificity of binding.

Such a partially non-random primer (hereafter called the partly fixed primer, or partly non-random primer, meaning that part of its sequence is fixed) can "anneal" to only the sequence at which the fixed sequence exists. That is, from the first primer, the partly fixed primer will bind at an average distance of 1024 characters (for a fixed five nucleotide characters). This primer will bind specifically only at the location of the occurrence of the particular five nucleotide sequence with respect to the first primer. The average distance between the first primer and the second non-random primer is ideal for DNA amplification and DNA sequencing. In this situation, the first primer is labeled. Thus, although there would be many locations in the long DNA molecule at which the non-random primer can bind, it would not affect the DNA sequencing because it is dependent only upon the labeled primer.

Although the partly fixed second primer has a random sequence component in it, a sub-population of the primer molecules will have the exact sequence that would bind with the exact target sequence. The proportion of the molecules with exact sequence that would bind with the exact target sequence will vary depending on the number of random characters in the partly fixed second primer. For example, in a second primer 11 nucleotides long with 6 characters fixed and 5 characters random, one in ~1000 molecules will have the exact sequence complementary to the target sequence on the template. By increasing the concentration of the partly fixed second primer appropriately, a comfortable level of PCR amplification required for sequencing can be achieved. When primer concentration is increased, it requires an increase in the concentration of Magnesium, which is required for the function of the polymerase enzyme. The excess primers (and "primer-dimers" formed due to excess of primers) can be removed after amplification reaction by a gel-purification step.

Any non-specific binding by any population of the second primers to non-target sequences could be avoided by adjusting (increasing) the temperature of re-annealing appropriately during DNA amplification. It is well known that the change of even one nucleotide due to point-mutation in some cancer genes can be detected by DNA-hybridization. This technique is routinely used for diagnosing particular cancer genes (e.g. John Lyons, "Analysis of ras gene point mutations by PCR and oligonucleotide hybridization," in *PCR Protocols: A guide to methods and applications*, edited by Michael A Innis et al., (1990), Academic Press, New York). This is done by adjusting the "re-annealing" or "melting-temperature", and fine-tuning the reaction conditions. Thus the binding of non-specific sequences even with just one nucleotide difference compared to the target binding-site in the template sequence can be avoided.

It should also be noted that non-specific binding sites for the partly fixed second primers could be expected to occur statistically on a long genomic DNA at many places other than the target site which is close to the first primer. Amplification of non-specific DNA between these primer binding sites that could occur on opposite strands of the template DNA could happen. However, this would not affect the objective of the present invention of specific DNA sequencing of the target sequence. Because only the first primer is labeled radioactivity or fluorescently, only the reaction products of the target DNA will be visualized on the sequencing gel pattern. The presence of such non-specific amplification products in the reaction mixture will also not affect the DNA sequencing reaction.

Amplification of DNA will occur not only between the first primer and the partly fixed second primer that occurs closest downstream from the first primer, but also between the first primer and one or two subsequently occurring

second primers, depending upon the distance at which they occur. However, these amplification products will all start from the first primer and will proceed up to these second primers. Since the DNA sequencing products are visualized by labeling the first primer, and since the DNA synthesis during the sequencing reaction proceeds from the first primer, the presence of two or three amplification products that start from the first primer will not affect the DNA sequencing products and their visualization on gels. At the most, the intensity of the bands that are subsets of different amplification products will vary slightly on the gel, but not affect the gel pattern. In fact, it is expected that this phenomenon will enable the sequencing of a longer DNA strand where the closest downstream primer is too close to the first primer—thereby avoiding the need for sequencing from the first primer again using another partly fixed second primer.

The minimum length of primer for highly specific amplification between primers on a template DNA is usually considered to be about 15 nucleotides. However, in the present invention, this length can be reduced by increasing the G/C content of the fixed sequence to 12–14 nucleotides.

In essence, the basic procedure of the present invention is fully viable and feasible, and any non-specificity can be avoided by fine-tuning the reaction conditions such as adjusting the annealing temperature and reaction temperature during amplification, and/or adjusting the length and G/C content of the primers, which are routinely done in the standard PCR amplification protocol.

The primary advantage of the present invention is to provide an extremely specific second primer that would bind precisely to a sequence at an appropriate distance from the first primer resulting in the ability to sequence a DNA without the prior knowledge of the second primer. From the newly worked out DNA sequence, a primer sequence can be made complementary to a sequence located close to the downstream end. This can be used as the first primer in the next DNA amplification-sequencing reaction, and the unknown sequence downstream from it can be obtained by again using the same partly fixed primer that was used in the first round of sequencing as the second primer. Thus, knowing only one short sequence in a contiguously long DNA molecule, the entire sequence can be worked out using the present invention.

When the length of the fixed sequence in the partly fixed second primer is increased in the present invention, the distance from the first primer at which the second primer will bind on the template will also be correspondingly increased. For a 6 nucleotide fixed sequence, the median length of DNA amplified will be ~2800 nucleotides (mean 4096 nucleotides), and for a 7 nucleotide fixed sequence, the median length of amplified DNA will be ~11,000 nucleotides (mean=~16,000 nucleotides). However, even if the length of amplified DNA is several thousand nucleotides, still this DNA can be used in DNA sequencing procedures. Furthermore, the present invention can be used to amplify a DNA of length which is limited only by the inherent ability of PCR amplification. A technique known as "long PCR" is used to amplify long DNA sequences (Kainz et at., "In vitro amplification of DNA Fragments >10 kb," *Anal Biochem.*, 202:46 (1992); Ponce & Micol, "PCR amplification of long DNA fragments" *Nucleic Acids Research*, 20:623 (1992)).

Existing genome sequencing methods employ the breaking down of a very long genomic DNA into many small fragments, sub-cloning them, sequencing them, and then assembling the sequence of the long DNA. Typically, a genomic DNA is broken down and cloned into overlapping fragments of approx. one million nucleotides in "YAC" (Yeast Artificial Chromosome) clones, each YAC clone is again fragmented and sub-cloned into overlapping fragments of ~25,000 nucleotides in "cosmid" clones, and each cosmid clone in turn sub-cloned into overlapping fragments of ~1000 nucleotides in "M13 phage" or "plasmid" clones. These are sequenced randomly to assemble the larger sequences in the hierarchy. The present invention circumvents the need for breaking down and sub-cloning steps, making it greatly advantageous for contiguously sequencing long genomic DNA.

Extending the above invention, another invention is presented here. This extended invention would enable the sequencing of ~500 nucleotide long sequence somewhere within a given long DNA with no prior information of any sequence at all within the long DNA. The probability that any specific primer of length 10 nucleotides would occur somewhere in a DNA of about one million nucleotides is approximately 1. The probability that any primer of length 15 nucleotides occur somewhere in a genome of about one billion nucleotides is approximately 1. Thus, use of any exact primer of about 15 nucleotide sequence on a genomic DNA in the present invention as the first primer, and the use of the second partly fixed primer will enable the sequencing of the DNA sequence bracketed by the two primers somewhere in the genome. Thus, this procedure can be used to obtain an exact sequence of about 500 characters somewhere from a genome without the prior knowledge of any of its sequence at all. Thus, by using many different primers with arbitrary but exact sequences, one can obtain many ~500-nucleotide sequences at random locations within a genome. Using these sequences as the starting points for contiguous genome sequencing in the present invention, the whole genomic sequence can be closed and completed. Thus an advantage of the present invention is that without any prior knowledge of any sequence in a genome, the whole sequence of a genome can be obtained.

It must be noted that every 15-nucleotide arbitrary primer may not always have a complementary sequence in a genome (of ~one billion nucleotides long). However, most often it would be present and would be useful in performing the above-mentioned sequencing. In some cases, there may be more than one occurrence of the primer sequence in the genome, and so may not be useful in obtaining the sequence. However, the frequency of successful single-hits can be extremely high (~90%) and can be further refined by using an appropriate length of the arbitrary primer. For genomes (or long DNAs) that are shorter than a billion nucleotides, shorter exact sequences in the first primers (say 10 characters) could be used, and the rest could be random or "degenerate" nucleotides. While this primer will still bind at the sequence complementary to the exact sequence, the longer primer will aid in avoiding non-specific DNA amplification. The length of the first primer can thus be increased using degenerate nucleotides at the ends to a desired extent, without affecting any specificity. Once a sequence is known in an unknown genomic DNA, then the present method can be performed to extend a contiguous sequence in both directions of the DNA from this starting point.

The present invention can also be useful to amplify the DNA between the first primer and the partly fixed second primer, with an aim to using this amplified DNA for purposes other than DNA sequencing, such as cloning. Although there would be sufficient quantity of the target specific amplified DNA in the reaction products, the reaction products will, however, contain the population of non-specific DNA amplified between the non-specifically occur-

ring second primer binding sites on opposite strands. However, by introducing a purification step from this reaction mixture, such as using an immobilized column containing only the first primer, the amplified target DNA can be purified and used for any other purposes.

UTILITY OF THE INVENTION

The present invention enables the amplification of a DNA adjacent to a known sequence using the PCR, without the knowledge of the sequence for a second primer.

The present primary invention provides a new method for sequencing a contiguously very long DNA sequence using the PCR technique, thereby enabling contiguous genomic sequencing. It will avoid the need for mapping or subcloning of shorter DNA fragments from haploid genomes such as the bacterial genomes. This method can be used on very large DNA inserts into vectors such as the YAC. Thus, diploid genomes can be sequenced without any further need to sub-clone from the YAC clones. The cloned inserts can be of any length, of several million nucleotides. Alternatively, wherever purified chromosomes are available, this method can be directly applied to sequence the whole chromosome without any need to fragment the chromosome or obtain YAC clones from the chromosome. This method can also be used on whole unpurified genomes with appropriate modifications to account for the allelic variations of the two alleles present on the two chromosomes. In essence, using the method of the present invention, one can generate contiguous genomic sequence information in a manner not possible with any other known protocol using PCR.

The present invention can find applications in many fields, for instance, medical, diagnostic, forensic, genetics, biotechnology, and genome research. It should be noted that this technique would be applicable in many other fields and instances, and such applications would be discernible by people of ordinary skills in the respective fields.

The extended invention that enables the sequencing of an unknown region of very long DNA (e.g. genomic DNA) of totally unknown sequence would also find many applications in biology and medicine. For instance, it can be used to physically "map" a chromosome or genome. It would, for example, enable the production of an inventory of many ~500 nucleotide long sequences and the exact primer associated with each of them. This method would also enable the cloning of the amplified DNA sequences from arbitrary regions from a genomic DNA without the need for breaking down the DNA. Using appropriately longer partly fixed primers (as the second primers), very long DNA pieces (several kilobases long) could be amplified and cloned by using this method.

BRIEF DESCRIPTION OF THE DRAWINGS

FIGS. 1(A–C) shows that the partly fixed second primer can bind only to the sequence locations in the "template" DNA (the DNA to be sequenced) containing the complementary sequence to the specific fixed sequence. The fixed sequence in the second primer is long enough and the rest of the random sequence is short enough that the random binding of this primer to any other sequence location in the template DNA is automatically precluded. This ensures that the second primer will bind extremely specifically to the sequence complementary to the fixed four or five nucleotide sequence used in the second primer, under the reaction conditions generally used in the PCR protocols. It should be noted that any sequence of five nucleotide length could be used in the fixed primer and the rest being random sequence. This is because, statistically, any of these sequences would occur at an appropriate average distance from the first primer. This is one of the primary advantages offered by the new invention. The advantage of this method stems from the chance occurrences of fixed sequences of particular length appropriate for DNA sequencing by the PCR technique. The statistical distribution of sequences of particular lengths has been worked out by Senapathy. Senapathy has also shown that natural DNA is essentially random in DNA sequence (P. Senapathy, "Origin of Eukaryotic Introns: A hypothesis, based on codon distribution statistics in genes, and its implications," *Proceedings of the National Academy of Sciences*, USA. 83:2133–2137 (1986)). The fixed sequence in the partly fixed second primer can be present at either ends of the primer or anywhere within the primer. Also, the fixed sequence within the second primer can be of any length, 4, 5, 6, or longer oligonucleotides. Furthermore, the fixed sequence can be split into 2 or 3 shorter fixed sequences at various positions within the second primer, still giving the same statistical result and binding property.

FIG. 2 shows that the mean-length between the first primer and the partly fixed second primer is appropriate for DNA sequencing using the PCR technique. For instance, the median length at which the sequence complementary to the second primer will be found is 730 nucleotides, which is ideal for DNA sequencing. This means, although the second primer will occur anywhere from zero distance up to about 6000 nucleotides, 50% of the time it will occur at around 730 nucleotides. Thus, by using two different fixed five nucleotide sequences in two different second primers, the chance that the DNA sequence of appropriate length can be sequenced is achieved with a probability of 99.9%. In the actual protocol, the length of a DNA sequence obtained in the new method would be known only when the results of a sequencing experiment is obtained. At this stage, if sufficient length of sequence is not obtained, then another second primer with a different fixed sequence could be used. With only a few different fixed five or six nucleotide sequences in the second primers, contiguous genomic sequencing can be systematically carried out. This means one needs to prepare only a few, say ten different second primers for sequencing the whole genome, which can be done at relatively very low expense and in bulk at one time at the start of sequencing a genome.

FIG. 3 shows how the partly fixed primer is used as the unknown second primer in contiguous genome sequencing. In a long template DNA, a sequence at the starting position should be known from which a first primer could be made. From this point, a streteh of the DNA sequence can be obtained using the new method. An appropriate sequence is selected from the downstream end of this sequence for making a primer that will be used as the first known primer for extending the sequencing. Using this primer and the same partly fixed primer as the unknown second primer, the sequence is extended further. This procedure is continuously repeated until the end of the sequence is reached.

FIG. 4 is a schematic indicating that the present invention enables the sequencing of a very long DNA in both directions from a starting known sequence location. From a known short sequence of only about a hundred nucleotides, two primers can be prepared such that they bind to opposite strands of the DNA. Using each of these known primers and the same second, partly fixed primer, sequencing can be extended in opposite directions on the DNA from the starting location.

FIGS. 5(A–B) describes the method to obtain the sequence of about 500 nucleotides from a genome or a very

long DNA, from which absolutely no sequence information is available. Depending on the length of the very long DNA (or the genome), a primer with an arbitrary but exact sequence is designed such that it would have approximately one binding site in the long DNA. This primer binding site will also have a site close to it (at an average distance of about 800 nucleotides) that will bind with the second, partly fixed (5 nucleotide) sequence primer. With the first primer radiolabeled or fluorescent labeled, the DNA sequence between the two primers can be obtained by performing PCR amplification and DNA sequencing. It should be noted that this is possible only because the unknown second, partly fixed primer will almost certainly occur within a distance ideal for PCR amplification and DNA sequencing from the first primer—no matter where in the long DNA the first primer occurs.

DETAILED DESCRIPTION OF THE INVENTION

The above and various other objects and advantages of the present invention are achieved by a method comprising:

a) synthesizing a partly fixed primer, with 4, 5, 6 nucleotide, or longer sequence characters fixed within it. The fixed sequence can be any sequence, with some preferred sequences such as those containing many G–C pairs that increases binding affinity. The fixed position within the primer can be anywhere, with some preferred positions;

b) taking a very long genomic DNA, either uncloned or a cloned large insert such as the YAC or cosmid in which a short sequence of about 20 characters somewhere within the DNA is known;

c) synthesizing a primer from the sequence known from the DNA in step b;

d) radiolabeling the primer in step c;

e) annealing the primers (from step a, and step d or step g as appropriate) to the DNA in step b, and amplifying the DNA between the attached primers;

f) performing DNA sequencing of the amplified DNA by the chemical degradation method of Maxam and Gilbert, or carrying out DNA sequencing by the Sanger method, or by modified PCR-sequencing method;

g) after obtaining the DNA sequence from step f, selecting an appropriate first primer towards the 3' end of the sequence, synthesizing it, and radiolabeling it;

h) repeating the steps e through g with the two primers (the same partly fixed unknown primer as the second primer and the newly synthesized primer from step g as the first primer);

i) if the sequence obtained in step f is too short to be of value, using another partly fixed primer with a different fixed sequence and the same first primer to obtain a longer DNA sequence.

Unless defined otherwise, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this invention belongs. All publications mentioned hereunder are incorporated herein by reference. Unless mentioned otherwise, the techniques employed herein are standard methodologies well known to one of ordinary skill in the art.

The partly fixed primer used to perform DNA amplification and sequencing are, of course, not limited to those described under the examples. Further modification in the method may be made by varying the length, content and position of the fixed sequence and the length of the random sequence. Additional obvious modifications include using different DNA polymerases and altering the reaction conditions of DNA amplification and DNA sequencing. Furthermore, the basic technique can be used for sequencing RNA using appropriate enzymes.

Instead of preparing the first primer completely, it can also be prepared as follows. Two or three shorter oligonucleotides that would comprise the complete primer could be ligated, by joining end-to-end after annealing to the template DNA, as described under another patent (Helmut Blocker, U.S. Pat. No. 5,114,839, 435/6, 5/1992) or as described in the publication (L. E. Kotler, et al., *Proceedings of the National Academy of Science, USA*, 90:4241–4245 (1993)). Alternatively, it can be synthesized using the single-stranded DNA binding protein, the subject of another invention (J. Kieleczawa, et al., *Science*, 258:1787–1791 (1992)). One of such procedures, or an improved version thereof, can be used to make the first primer in the present invention. All in all, the first primer need not be synthesized at every PCR reaction while contiguously sequencing a long DNA, and can be directly constructed from an oligonucleotide bank. Based on the present invention, the second primer also can be chosen from a set of only a few pre-prepared primers. This enables the direct automation of sequencing the whole long DNA by incorporating the primer elements into the series of sequential PCR reactions.

An advantage of the present invention is that from a known sequence in a very long DNA, sequencing can be performed in both directions on the DNA. Two first primers can be prepared, one on each strand, running in the opposite directions, and the sequence can be extended on both directions until the two very ends of the long DNA are reached by the present invention, using a small set of pre-prepared partly fixed second primers.

One of the major advantages of the present invention is that it is highly amenable to various kinds of automation. Instead of radiolabeling the first known primer, it can be fluorescently labeled, and with this the DNA sequencing can be performed in an automated procedure on machines such as that marketed by the Applied Biosystems ("373 DNA Sequencer: Automated sequencing, sizing, and quantitation", a pamphlet from the Applied Biosystems, A Division of Perkin-Elmer Corporation (1994)). In the present invention there is no need to newly synthesize any primers to sequence a very long DNA. Thus, with the pre-prepared set of partly fixed second primers, an oligonucleotide bank for the synthesis of the first primer, and a large supply of the template genomic DNA (or any long DNA), the sequencing of the whole long DNA can be automated using robots almost without any human intervention, except for changing the sequencing gels.

The following processes can be computer controlled: 1) the selection of the appropriate sequence for constructing the first primer close to the 3' end of the newly worked out sequence, 2) determining whether the sequence obtained is too short and selection of a different partly fixed second primer, 3) assembling the contiguous DNA sequences from the various lanes and various gels and appending to a database, and other such processes. Thus the present invention enables the construction of a fully automated contiguous DNA sequencing system. Any such automations are obvious modifications to the present invention.

The present invention is not limited to only unknown genomic DNA, and can be used to sequence any DNA under any situations. DNAs or RNAs of many different origins (e.g. viral, cDNA, mRNA) can be sequenced not only limited to research or information gathering purposes, but

also to other purposes such as disease diagnosis and treatment, DNA testing, and forensic applications.

It is understood that the examples and embodiments described herein are for illustrative purposes only and that various modifications or changes in light thereof will be suggested to persons skilled in the art and are to be included within the spirit and purview of this application and scope of the appended claims.

It should be noted that any kit or process used for research, diagnostic, forensic, treatment, production or other purposes that uses the present invention is covered under these claims. Furthermore, the various sequences of the partly fixed second primers that can be used in the present invention are covered under this patent. Thus, any kit or process that uses this method and/or the DNA strands with the sequences that would comprise the partly fixed second primers will also be covered under this.

In addition to contiguous DNA sequencing, the present invention will cover the amplification of the DNA strands that are bounded between the known primer and the partly fixed second primer (either from claim 1 or from claim 2). The DNA amplification can also be performed for long DNA strands using the long PCR amplification protocols.

What is claimed is:

1. A method for amplifying by the polymerase chain reaction (PCR) a portion of a target nucleic acid sequence comprising the following steps:

a. providing a first primer wherein the first primer consists of a sequence fully complementary to a first primer binding site on said target nucleic acid sequence;

b. providing a second primer wherein said second primer consists of about 12–16 nucleotides of which 1–10 nucleotides anywhere within it are of fixed sequence, while the remaining nucleotides of the second primer are of random sequence, wherein the second primer is fully complementary to a second primer binding site; and

c. performing PCR amplification using the primers of (a) and (b) and said target nucleic acid to effect the amplification of a portion of said target nucleic acid, wherein said amplification is under conditions of sufficient stringency so that specific amplification occurs.

2. The method of claim 1 further comprising sequencing the amplification product of step (c) thereby determining the sequence of a portion of said target nucleic acid sequence.

3. The method of claim 1 wherein the target nucleic acid is DNA.

4. The method of claim 1 wherein the target nucleic acid is RNA.

5. The method of claim 2 wherein the target nucleic acid is DNA.

6. The method of claim 2 wherein the target nucleic acid is RNA.

* * * * *

102RFB

Print Request: LEXSEE

Time of Request: June 11, 2001 12:16 pm EST

Number of Lines: 416
Job Number: 59:0:29065784

Client ID/Project Name:

Research Information:

Lexsee 108 F.3d 1361

Note:

LEXSEE 108 F.3d 1361

GENENTECH, INC., Plaintiff-Appellee, v. NOVO NORDISK A/S, NOVO NORDISK OF NORTH AMERICA, INC., and NOVO NORDISK PHARMACEUTICALS, INC., Defendants-Appellants.

96-1440

UNITED STATES COURT OF APPEALS FOR THE FEDERAL CIRCUIT

108 F.3d 1361; 1997 U.S. App. LEXIS 4808; 42 U.S.P.Q.2D (BNA) 1001

March 13, 1997, Decided

SUBSEQUENT HISTORY:
[**1]

Rehearing Denied and In Banc Suggestion Declined May 12, 1997, Reported at: 1997 U.S. App. LEXIS 15056. Certiorari Denied November 3, 1997, Reported at: 1997 U.S. LEXIS 6493.

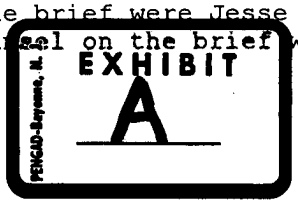
PRIOR HISTORY:
Appealed from: U.S. District Court for the Southern District of New York. Senior Judge Motley.

DISPOSITION:
VACATED.

COUNSEL:
NOVO NORDISK, A/S, NOVO NORDISK OF NORTH AMERICA, INC. and NOVO NORDISK PHARMACEUTICALS, INC., Defendants-Appellants.

Leora Ben-Ami, Rogers & Wells, of New York, New York, argued for plaintiff-appellee. With her on the brief were John E. Kidd, Nicholas L. Coch, Joseph Ferraro, Philip E. Roux, and Gerard P. Norton. Of counsel was Ryan Trainer, Rogers & Wells, of Washington, D.C.

Albert L. Jacobs, Jr., Graham & James LLP, of New York, New York, argued for defendant-appellants. With him on the brief were Jesse D. Reingold, Gerard F. Diebner, and Daniel A. Ladow. Of counsel on the brief were John C. Vassil, Kurt



E. Richter and Kenneth H. Sonnenfeld, Morgan & Finnegan, LLP, of New York, New York. Of counsel were Brad S. Needleman and Andrew T. Solomon, Graham & James.

JUDGES:
Before ARCHER, Chief Judge, LOURIE, and BRYSON, Circuit Judges.

OPINIONBY:
LOURIE

OPINION:

[*1362] LOURIE, Circuit Judge.

Novo Nordisk A/S, Novo Nordisk of North America, Inc., and Novo Nordisk Pharmaceuticals, Inc. (collectively "Novo") appeal from the order [**2] of the United States District Court for the Southern District of New York, issuing a preliminary injunction in favor of Genentech, Inc., enjoining Novo from importing, marketing, using, selling, offering for sale or distributing its Norditropin (R) -brand recombinant human growth hormone (hGH) [*1363] product. Genentech, Inc. v. Novo Nordisk A/S, 935 F. Supp. 260 (S.D.N.Y. 1996). Because the district court's conclusion that Genentech had demonstrated a likelihood of success on the merits was based on an error of law and because its remaining findings were premised on this error, we vacate the injunction.

BACKGROUND

This consolidated patent infringement action was first brought in the United States District Court for the Southern District of New York on November 30, 1994. On May 12, 1995, Genentech moved for a preliminary injunction under U.S. Patent 4,601,980 to prevent Novo from importing, marketing, using, selling, offering for sale or distributing in the United States its Norditropin (R) -brand recombinant hGH product. The district court granted Genentech's motion and issued an injunction. Novo Nordisk of North Am., Inc. v. Genentech, Inc., 1995 U.S. Dist. LEXIS 12588, No. 94 Civ. 8634 (CBM), [**3] 1995 WL 512171 (S.D.N.Y. Aug. 28, 1995).

On appeal this court vacated the injunction. Novo Nordisk of North Am., Inc. v. Genentech, Inc., 77 F.3d 1364, 37 U.S.P.Q.2D (BNA) 1773 (Fed. Cir. 1996). We held that the district court clearly erred in finding that Genentech established a likelihood of proving infringement of the '980 patent because that finding was based on an improper construction of claim 2 of the patent. Based upon the specification and prosecution history, we concluded that because the claim used the phrase "human growth hormone unaccompanied by ... other extraneous protein," it was limited to processes for directly expressing either hGH or met-hGH. Id. at 1371, 37 U.S.P.Q.2D (BNA) at 1779. Because the parties agreed that Novo did not use direct expression to produce these proteins, we concluded that Novo did not infringe the patent. Id.

Upon returning to the district court, Genentech asserted its newly issued U.S. Patent 5,424,199. The '199 patent has the same specification as the '980 patent and contains a single claim directed to:

[a] method of producing a protein consisting essentially of amino acids 1-191 of human growth hormone [**4] comprising:

(a) expressing in a transformant bacterium, DNA coding for a human growth hormone conjugate protein, which conjugate protein consists essentially of amino acids 1-191 of human growth hormone as set forth in combined Figs. 1 and 3 unaccompanied by the leader sequence of human growth hormone or other extraneous protein bound thereto and an additional amino acid sequence which is specifically cleavable by enzymatic action, and

(b) cleaving extracellularly said conjugate protein by enzymatic action to produce said protein consisting essentially of amino acids 1-191 of human growth hormone.

This claim differs from the claim adjudicated in the prior case in reciting that

the encoded protein has an additional amino acid sequence and includes the step of cleaving this conjugate protein. This process of expressing a DNA encoding a conjugate protein and using an enzyme to cleave off an undesired portion of that protein is generally known as cleavable fusion expression. The parties agree that Novo uses cleavable fusion expression to produce hGH. Id.

On June 27, 1996, after conducting a twelve-day evidentiary hearing, the district court again issued a preliminary [**5] injunction, this time based upon the '199 patent, enjoining Novo from importing, marketing, using, selling, offering for sale, or distributing in the United States its Norditropin (R) -brand recombinant hGH product. Genentech v. Novo Nordisk A/S, 935 F. Supp. 260 (S.D.N.Y. 1996). The district court based its decision upon, inter alia, a finding that Genentech would likely overcome Novo's defense that the '199 patent was invalid for lack of an enabling disclosure under 35 U.S.C. @ 112, P 1 (1994).

Novo appeals to this court, challenging the grant of the preliminary injunction. n1 We [*1364] have jurisdiction pursuant to 28 U.S.C. @ 1292(c) (1994).

n1 On July 3, Novo moved for an emergency stay of the injunction pending disposition of this appeal. On August 1, we denied Novo's motion and reinstated the injunction. However, after having heard oral argument in this case, we reconsidered the motion and reinstated the stay of the injunction.

DISCUSSION

The grant or denial of a preliminary injunction pursuant [**6] to 35 U.S.C. @ 283 is within the discretion of a district court. We Care, Inc. v. Ultra-Mark Int'l Corp., 930 F.2d 1567, 1570, 18 U.S.P.Q.2D (BNA) 1562, 1564 (Fed. Cir. 1991). Accordingly, a trial court's decision granting a preliminary injunction will be overturned on appeal only upon a showing that the court abused its discretion. Joy Techs., Inc. v. Flakt, Inc., 6 F.3d 770, 772, 28 U.S.P.Q.2D (BNA) 1378, 1380 (Fed. Cir. 1993). Such an abuse of discretion may be established by showing that the court made a clear error of judgment in weighing relevant factors or exercised its discretion based upon an error of law or clearly erroneous factual findings. Id.

As the moving party, Genentech had to establish its right to a preliminary injunction in light of four factors: (1) a reasonable likelihood of success on the merits; (2) irreparable harm if the injunction were not granted; (3) the balance of the hardships; and (4) the impact of the injunction on the public interest. Nutrition 21 v. United States, 930 F.2d 867, 869, 18 U.S.P.Q.2D (BNA) 1347, 1348-49 (Fed. Cir. 1991); Hybritech Inc. v. Abbott Lab., 849 F.2d 1446, 1451, 7 U.S.P.Q.2D (BNA) 1191, 1195 (Fed. Cir. 1988).

A. Likelihood of Success [**7] on the Merits

In order to demonstrate that it has a likelihood of success, Genentech must show that, in light of the presumptions and burdens that will inhere at trial on the merits, (1) it will likely prove that Novo infringes the '199 patent and (2) its infringement claim will likely withstand Novo's challenges to the validity and enforceability of the '199 patent. See New England Braiding Co. v. A.W. Chesterton Co., 970 F.2d 878, 882-83, 23 U.S.P.Q.2D (BNA) 1622, 1625-26 (Fed.

Cir. 1992). n2 In other words, if Novo raises a "substantial question" concerning validity, enforceability, or infringement (i.e., asserts a defense that Genentech cannot show "lacks substantial merit") the preliminary injunction should not issue. Id. More specifically, with regard to Novo's validity defenses, the question on appeal is whether there is substantial merit to Novo's assertion that the '199 patent claim fails to meet the requirements of 35 U.S.C. @ 112, P 1 (1994).

n2 A patent is presumed valid, 35 U.S.C. @ 282 (1994), and a party challenging validity must prove invalidity by clear and convincing evidence. "However, the presumption does not relieve a patentee who moves for preliminary injunction from carrying the normal burden of demonstrating that it will likely succeed on all disputed liability issues at trial, even when the issue concerns the patent's validity." New England Braiding, 970 F.2d at 882, 23 U.S.P.Q.2D (BNA) at 1625 (citing Nutrition 21, 930 F.2d at 869, 18 U.S.P.Q.2D (BNA) at 1349).

[**8]

Novo argues that the district court's findings regarding validity under @ 112, P 1, are clearly erroneous because it presented clear and convincing evidence that the patent specification would not have enabled a person of ordinary skill in the art to practice the claimed invention without undue experimentation. Novo also argues that the specification fails to contain a written description of the claimed invention. Regarding enablement, Novo argues that the patent is invalid because it does not contain sufficient detail concerning the practice of the claimed method. Novo argues that the mere generic statement of the possibility of cleavable fusion expression, along with the DNA sequence encoding hGH, a single enzyme (trypsin) for cleaving undisclosed conjugate proteins, and a statement of that enzyme's cleavage sites as being potential amino acid extensions conjugated to hGH is not an enabling disclosure commensurate in scope with the claim. Genentech responds that all of the district court's factual findings regarding enablement are supported by the record. More specifically, Genentech argues that those skilled in the art of recombinant protein expression and purification at the [**9] time of filing, July 5, 1979, would have been able to use cleavable fusion expression to produce hGH without undue experimentation by using the teachings of the specification along with methods and tools well known in the art. We conclude that Novo has raised more [*1365] than a substantial question concerning the validity of the '199 patent. In fact, it has shown that the patent is invalid.

Section @ 112, P 1, provides, in relevant part that:

the specification shall contain a written description of the invention, and the manner and process of making and using it, in such full, clear, concise, and exact terms as to enable any person skilled in the art to which it pertains, or with which it is most nearly connected, to make and use the same ....

"To be enabling, the specification of a patent must teach those skilled in the art how to make and use the full scope of the claimed invention without 'undue experimentation.'" In re Wright, 999 F.2d 1557, 1561, 27 U.S.P.Q.2D (BNA) 1510, 1513 (Fed. Cir. 1993); see also Amgen Inc. v. Chugai Pharms. Co., 927 F.2d 1200, 1212, 18 U.S.P.Q.2D (BNA) 1016, 1026 (Fed. Cir. 1991); In re Fisher, 57 C.C.P.A.

1099, 427 F.2d 833, 839, 166 U.S.P.Q. (BNA) 18, 24 (CCPA 1970) ("The scope [**10] of the claims must bear a reasonable correlation to the scope of enablement provided by the specification to persons of ordinary skill in the art."). Whether making and using the invention would have required undue

experimentation, and thus whether the disclosure is enabling, is a legal conclusion based upon several underlying factual inquiries. See In re Wands, 858 F.2d 731, 735, 736-37, 8 U.S.P.Q.2D (BNA) 1400, 1402, 1404 (Fed. Cir. 1988).

The question before us is whether the specification would have enabled a person having ordinary skill in the art at the time of filing to use cleavable fusion expression to make hGH without undue experimentation. There is no dispute that the portion of the specification chiefly relied upon by Genentech and by the district court, column 7, lines 29-59, does not describe in any detail whatsoever how to make hGH using cleavable fusion expression. For example, no reaction conditions for the steps needed to produce hGH are provided; no description of any specific cleavable conjugate protein appears. The relevant portion of the specification merely describes three (or perhaps four) applications for which cleavable fusion expression is generally well-suited [**11] and then names an enzyme that might be used as a cleavage agent (trypsin), along with sites at which it cleaves ("arg-arg or lys-lys, etc."). n3 Thus, the specification does not describe a specific material to be cleaved or any reaction conditions under which cleavable fusion expression would work.

n3 At column 7, lines 52-58, the specification states: "At least in the latter three applications [of the four applications that are disclosed], the synthetic adaptor molecular [sic] employed to complete the coding sequence of the mRNA transcript can additionally incorporate codons for amino acid sequences specifically cleavable, as by enzymatic action. For example, trypsin will cleave specifically at arg-arg or lys-lys, etc."

Notwithstanding this limited disclosure, Genentech argues (and the district court found) that those of ordinary skill in the art would have been able to practice the claimed invention without undue experimentation. Essentially, Genentech's argument is that the knowledge of one skilled in [**12] the art was sufficient to provide all of the missing information and, more specifically, that the disclosure of a DNA encoding hGH, when combined with prior art cleavable fusion expression techniques applied to non-human proteins, would enable the practice of the claimed method. In support of this argument, Genentech points to the testimony of Dr. Ravetch, who testified as to the knowledge of one skilled in the art, to the extensive description of enzymes in the reference textbook Methods in Enzymology, and to the specification's explicit reference to British Patent 2008123-A, which more fully details the potential use of trypsin in cleavable fusion expression.

In response to these arguments, Novo asserts that at the time of filing, trypsin and other like enzymes were used only to digest proteins, not to specifically and precisely cleave conjugate proteins to yield intact, useful proteins, and that the British patent explicitly indicates that trypsin would not be useful for the cleavable fusion expression of arginine-containing proteins such as hGH. Novo further argues that neither the specification nor the references cited by Genentech suggest a single amino acid sequence, out [**13] of the virtually infinite range of possibilities, [*1366] that would yield hGH

108 F.3d 1361, *1366; 1997 U.S. App. LEXIS 4808, **13;
42 U.S.P.Q.2D (BNA) 1001

in a useful form when cleaved from the conjugate protein.

We agree with Novo. Genentech's arguments, focused almost exclusively on the level of skill in the art, ignore the essence of the enablement requirement. Patent protection is granted in return for an enabling disclosure of an invention, not for vague intimations of general ideas that may or may not be workable. See Brenner v. Manson, 383 U.S. 519, 536, 148 U.S.P.Q. (BNA) 689, 696, 16 L. Ed. 2d 69, 86 S. Ct. 1033 (1966) (stating, in context of the utility requirement, that "a patent is not a hunting license. It is not a reward for the search, but compensation for its successful conclusion.") Tossing out the mere germ of an idea does not constitute enabling disclosure. While every aspect of a

generic claim certainly need not have been carried out by an inventor, or exemplified in the specification, reasonable detail must be provided in order to enable members of the public to understand and carry out the invention. That requirement has not been met in this specification with respect to the cleavable fusion expression of hGH.

It is true, as Genentech argues, that a specification need not disclose [**14] what is well known in the art. See, e.g., Hybritech Inc. v. Monoclonal Antibodies, Inc., 802 F.2d 1367, 1385, 231 U.S.P.Q. (BNA) 81, 94 (Fed. Cir. 1986). However, that general, oft-repeated statement is merely a rule of supplementation, not a substitute for a basic enabling disclosure. It means that the omission of minor details does not cause a specification to fail to meet the enablement requirement. However, when there is no disclosure of any specific starting material or of any of the conditions under which a process can be carried out, undue experimentation is required; there is a failure to meet the enablement requirement that cannot be rectified by asserting that all the disclosure related to the process is within the skill of the art. It is the specification, not the knowledge of one skilled in the art, that must supply the novel aspects of an invention in order to constitute adequate enablement. This specification provides only a starting point, a direction for further research.

The specification indicates that it purports to solve a problem. That problem is summarized at column 3, line 65, through column 4, line 8:

[A] need has existed for new methods of producing [**15] hGH and other polypeptide products in quantity and that need has been particularly acute in the case of polypeptides too large to admit to organic synthesis or, for that matter, microbial expression from entirely synthetic genes. Expression of mammalian hormones from mRNA transcripts ... has permitted only microbial production of bio-inactive conjugates from which the desired hormone could not practically be cleaved.

The problem thus was the difficulty of obtaining hGH from a precursor containing added protein material. This problem was solved by the description of a method of obtaining hGH unaccompanied by a leader sequence or other extraneous proteins, as claimed in the '980 patent. However, the specification for the '199 patent, which is the same as the specification for the '980 patent, does not provide a specific enabling disclosure concerning what the new claim recites, viz., obtaining hGH by cleaving an hGH-containing conjugate protein. That was the problem avoided by the invention claimed in the '980 patent. The present specification contains no more disclosure than the '980 specification, but this patent now purports to claim the unresolved problem that the '980 [**16] patent overcame. Genentech is attempting to bootstrap a vague statement of a problem

into an enabling disclosure sufficient to dominate someone else's solution of the problem. This it cannot do.

Genentech's arguments in favor of enablement are unavailing. While Genentech's witness, Dr. Ravetch, did state that it would have been possible for a skilled artisan to create a DNA sequence coding for arg-arg-hGH or lys-lys-hGH, he did not discuss the experimentation needed for the creation of DNA coding for more extensive sequences, such as those that have proved necessary to the production of hGH via cleavable fusion expression. Likewise, the description of a wide range of enzymes in Methods in Enzymology, by itself, does not render routine the [*1367] determination of an enzyme-conjugate protein combination. Rather, as Novo argues and the record reflects, various combinations of conjugate protein sequences, cleaving enzymes, and reaction conditions needed to be studied to establish a process for producing hGH in useful form. Finally, the British patent cited in the specification actually works against Genentech's position by explicitly teaching that trypsin would not work well to produce [**17] hGH. The specification does not even acknowledge any of the known difficulties associated with using trypsin on an hGH conjugate

protein. This specification is so lacking with respect to the limitation of paragraph (b) of claim 1 that providing testimony regarding the skill in the art has been an exercise in futility.

The limited testimony regarding the knowledge of one skilled in the art offered by Genentech at the preliminary injunction hearing, and relied upon by the district court, is further undermined by the fact that no one had been able to produce any human protein via cleavable fusion expression as of the application date. If, as Genentech argues, one skilled in the art, armed only with what the patent specification discloses (a DNA sequence encoding a human protein, in this case, hGH, and a single example of an enzyme and its cleavage site), could have used cleavable fusion expression to make a human protein without undue experimentation, it is remarkable that this method was not used to make any human protein for nearly a year, see Shine et al., 285 Nature 456 (June 1980), or to make hGH for five years. See Belagaje et al., 3 DNA 120 (1984). Certainly, [**18] DNAs encoding desirable human proteins were known at the time of filing (e.g., insulin, described in the British patent), and a great many researchers were attempting to produce human proteins using recombinant DNA technology. This failure of skilled scientists, who were supplied with the teachings that Genentech asserts were sufficient and who were clearly motivated to produce human proteins, indicates that producing hGH via cleavable fusion expression was not then within the skill of the art. The contrary testimony offered by Genentech's witnesses, who hypothesized about the skill of the art more than fifteen years earlier, does not demonstrate the incorrectness of Novo's arguments. See In re Buchner, 929 F.2d 660, 661, 18 U.S.P.Q.2D (BNA) 1331, 1332 (Fed. Cir. 1991) ("An expert's opinion on the ultimate legal issue [of enablement] must be supported by something more than a conclusory statement.").

Moreover, it stands to reason that if the disclosure of a useful conjugate protein and the method for its cleavage were so clearly within the skill of the art, it would have been expressly disclosed in the specification, and in the usual detail. Patent draftsmen are not loath to provide [**19] actual or constructive examples, with details, concerning how to make what they wish to claim. In addition, as indicated above, the specification of this patent was clearly drafted to claim the invention of obtaining hGH unaccompanied by

extraneous protein, the cleavage of which was identified by the specification as a problem in this field. Genentech's inventors knew how to enable that which they had invented. These facts underline the inadequacy of the specification in enabling that which it provided only a means to avoid.

The record does not support the district court's implicit finding that the disclosure of trypsin and its cleavage site enables the production of any conjugate protein from which hGH can practically be cleaved and thus produced in useful form; the record indicates that determination of these features required further undue experimentation. None of the expert testimony relied upon by Genentech or by the district court suggests otherwise. n4 Where, as here, the claimed invention is the application of an unpredictable technology in the early stages of development, an enabling [*1368] description in the specification must provide those skilled in the art with a specific [**20] and useful teaching. Genentech has not shown that the '199 patent provides that teaching.

n4 Novo's witness, Dr. Villa-Komaroff, merely stated on cross-examination that, assuming arg-arg-hGH was initially produced and successfully extracted from the transformed cell, that "under the best condition, approximately five percent of the time there will be in the [post-digestion] mix [hGH]." This statement, characterized by Genentech as an admission, was made in the limited context of partial trypsin digests of isolated arg-arg-hGH, but none of the necessary experimentation is described in the specification, which is where it should be if it is to contribute to an enabling disclosure.

Under the circumstances, we are compelled to conclude that the district court made an error of law in ruling that Genentech showed a likelihood of success on enablement. See In re Epstein, 32 F.3d 1559, 1568, 31 U.S.P.Q.2D (BNA) 1817, 1823 (Fed. Cir. 1994) ("Enablement is a question of law ... which may involve subsidiary questions of [**21] fact."). Furthermore, since we are able to review the record and to read the specification, there is no reason why we should limit our decision here to reversing the grant of the preliminary injunction. Rather, because the parties agreed at oral argument that the enablement issue had been thoroughly ventilated by the extensive arguments before the district court and that court's extensive analysis, n5 we deem it appropriate to rule on the merits of Novo's defense of invalidity. See 28 U.S.C. @ 2106 (1994) ("The Supreme Court or any other court of appellate jurisdiction may ... direct the entry of such appropriate judgment, decree, or order, or require such further proceedings to be had as may be just under the circumstances."); Chicago Observer, Inc. v. City of Chicago, 929 F.2d 325, 329 (7th Cir. 1991) (reversing preliminary injunction and instructing district court to enter judgment in favor of defendant because the plaintiff "has not suggested that it holds more evidence it could offer at trial and we cannot imagine what additional evidence could aid its cause. Litigation is costly not only for the litigants but also for parties in other cases waiting in the queue for [**22] judicial attention. Once it becomes clear that additional proceedings are pointless, the court should bring the case to a close."). We therefore hold that claim 1 and hence the '199 patent are invalid as a matter of law for failure of the specification to enable the practice of the claimed method.

n5 Genentech stated that it would introduce new evidence at a full trial only

in response to new arguments and new defenses raised by Novo. Novo revealed that it had no intention of raising any new arguments or defenses, stating that the "full and complete record" on appeal gave this court "the benefit of everything it really needs" to reach ultimate issues of validity. Thus, considerations that would normally dictate that we limit our decision to reversing the grant of the preliminary injunction are not present. See University of Texas v. Camenisch, 451 U.S. 390, 395, 68 L. Ed. 2d 175, 101 S. Ct. 1830 (1981) (stating that it is generally inappropriate to render a final judgment on the merits at the preliminary injunction stage because "a preliminary injunction is customarily granted on the basis of procedures that are less formal and evidence that is less complete than in a trial on the merits.") (citations omitted)(emphasis added).

[**23]

Novo has also argued that the '199 patent is invalid for lack of a written description of the claimed invention and that it is not infringed by Novo. Givenour decision on the enablement question, we need not reach these issues.

B. Other Factors

Novo also challenges the district court's findings that irreparable harm, the equities, and the public interest favored Genentech. In view of our conclusion concerning the invalidity of the '199 patent, we need not consider these other findings.

CONCLUSION

The court abused its discretion by granting the preliminary injunction based upon an error of law. The district court's error was in finding that Genentech had shown a likelihood of success on the merits since the '199 patent is invalid for failure of the specification to meet the enablement requirement of @ 112,

P1. Accordingly, we vacate the injunction and instruct the district court to dismiss Genentech's claim for infringement of the '199 patent on the ground that the patent is invalid.

VACATED

102RFB

********** Print Completed **********

Time of Request: June 11, 2001 12:16 pm EST

Print Number: 59:0:29065784
Number of Lines: 416
Number of Pages: 9